Physics 403

Common Probability Distributions

Segev BenZvi

Department of Physics and Astronomy
University of Rochester

Table of Contents

Last Time

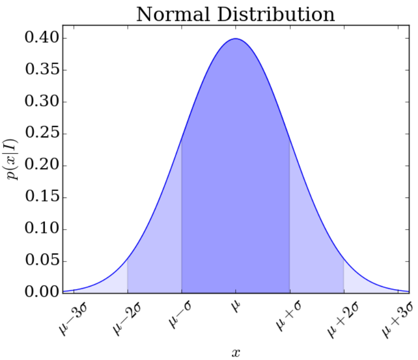
- Probability density functions
- Summary Statistics:
  - Location parameters: mean, median, mode
  - Width parameters: variance, covariance
  - Higher-order moments: skew, kurtosis
  - Ordered rank statistics: percentiles
  - The cumulative distribution function
  - Histograms

Last Time

The 68-95-99 Rule

In physics we tend to express rare events in terms of the tails of the Gaussian PDF

$$p(x|I) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$



The "68-95-99" quantile rule:

- 68.27% of the data are within 1σ of the mean.
- 95.45% of the data are within 2σ of the mean.
- 99.73% of the data are within 3σ of the mean.

Table of Contents

Reading for Today

- Cowan: Chapter 2
- *Numerical Recipes in C*: Chapter 7

Binomial Distribution

- Bernoulli trials — i.e, binary measurements which result in "success" with probability p and "failure" with probability $1 - p$ — are described by the binomial distribution.
- In n trials, like a coin toss, the probability of m "heads" is

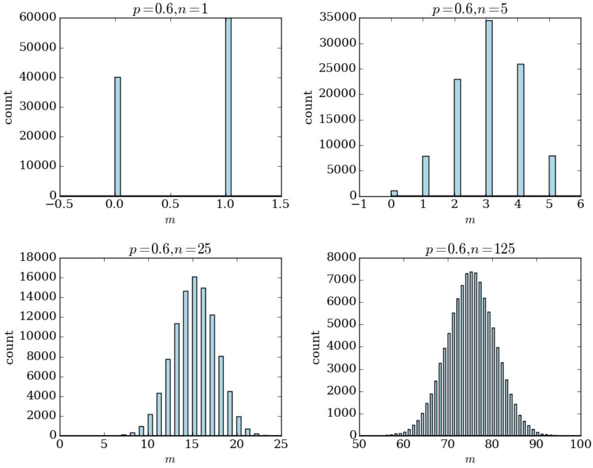
$$p^m(1-p)^{n-m}$$

- If we don't care about the order of the successes, then there are ${}_nC_m$ ways to get m successes in m trials. Therefore,

$$p(m|n,p) = \frac{n!}{m!(n-m)!}p^m(1-p)^{n-m}$$

Binomial Distribution

The binomial PDF is a discrete distribution:



Note how the binomial looks increasingly Gaussian as $n \to$ large.

Binomial Distribution

Mean

The mean of the binomial distribution is

$$\begin{aligned}
\langle m \rangle &= \sum_{m=0}^{n} m \cdot \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m} \\
&= np \sum_{m=1}^{n} \frac{(n-1)!}{(m-1)!(n-m)!} p^{m-1} (1-p)^{n-m} \\
&= np \sum_{m'=0}^{n'} \frac{n'!}{m'!(n'-m')!} p^{m'} (1-p)^{n'-m'} \\
&= np
\end{aligned}$$

where we simply used the fact that $p(m|n,p)$ is normalized over the sum from $m = 0$ to n.

Binomial Distribution

Variance

To find the variance $\mathrm{V}(m)$, note that

$$\begin{aligned}\langle m(m-1)\rangle &= \sum_{m=0}^{n} m(m-1)\cdot\frac{n!}{m!(n-m)!}p^m(1-p)^{n-m}\\ &= n(n-1)p^2\sum_{m'=0}^{n'}\frac{n'!}{m'!(n'-m')!}p^{m'}(1-p)^{n'-m'}\\ \langle m^2-m\rangle &= n(n-1)p^2\end{aligned}$$

where $m'=m-2$, $n'=n-2$, and the sum is 1. Therefore,

$$\begin{aligned}\mathrm{V}(m) &= \langle m^2\rangle - \langle m\rangle^2 = \langle m^2-m\rangle + \langle m\rangle - \langle m\rangle^2\\ &= n(n-1)p^2 + np - (np)^2\\ &= np(1-p)\end{aligned}$$

Binomial Distribution

Detector Efficiencies

Example

You measure the tracks of cosmic ray particles using a stack of silicon detectors which are 95% efficient. You decide that 3 points are needed to define a track. How efficient is a stack of 3 layers? What about 4, or 5?

Binomial Distribution

Detector Efficiencies

Example

You measure the tracks of cosmic ray particles using a stack of silicon detectors which are 95% efficient. You decide that 3 points are needed to define a track. How efficient is a stack of 3 layers? What about 4, or 5?

$$
\begin{aligned}
P(3|p=0.95, n=3) &= 0.95^3 = 0.857 \\
P(3+4|p=0.95, n=4) &= P(3|\ldots) + P(4|\ldots) \\
&= \frac{4!}{3!1!}0.95^3 0.05 + 0.95^4 = 0.986 \\
P(3+4+5|p=0.95, n=5) &= P(3|\ldots) + P(4|\ldots) + P(5|\ldots) \\
&= \frac{5!}{3!2!}0.95^3 0.05^2 + \frac{5!}{4!1!}0.95^4 0.05 + 0.95^5 \\
&= 0.999
\end{aligned}
$$

Multinomial Distribution

Generalization of the Binomial Distribution

- If instead of two outcomes we have k, we can generalize the binomial distribution to the multinomial distribution:

$$p(m_1, m_2, \ldots, m_k | n, p_1, p_2, \ldots, p_k) = \frac{n!}{\prod_i m_i!} \prod_{i=1}^{k} p_i^{m_i}$$

where

$$\sum_{i=1}^{k} p_i = 1, \qquad \sum_{i=1}^{k} m_i = n$$

- The multinomial is a joint probability distribution over the $\{m_i\}$.

Example

Example: binned data. If you sample trials from a PDF and bin the results, the predicted counts in each bin will follow a multinomial distribution.

Poisson Distribution

- The Poisson distribution is a limiting case of the binomial distribution ($n \to \infty$, $p \to 0$, $\langle m \rangle \to$ finite).
- It applies when we observe particular outcomes but without knowledge of the number of trials. For example:
  - Number of lightning strikes in a thunderstorm
  - Number of supernova explosions in the Galaxy per century
- Suppose that on average λ events are expected to occur in some interval of length T. I.e., the events occur at constant rate R such that $\lambda = RT$.
- If we split the interval up into n sections so that in each section we observe 0 or 1 events, the probability of observing an event in a section is $p = \lambda/n$, and the total number of events in the interval follows a binomial distribution:

$$p(m|p = \lambda/n, n) = \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m}$$

Poisson Distribution

Letting $n \to \infty$ we find that

$$p(m|p = \lambda/n, n) = \lim_{n\to\infty} \frac{n!}{m!(n-m)!} \left(\frac{\lambda}{n}\right)^m \left(1 - \frac{\lambda}{n}\right)^{n-m}$$

The factorials reduct to a power of n in the large n limit:

$$\lim_{n\to\infty} \frac{n!}{(n-m)!} = \lim_{n\to\infty} n(n-1)(n-2)\ldots(n-m+1) \to n^m$$

And we use the definition of the exponential:

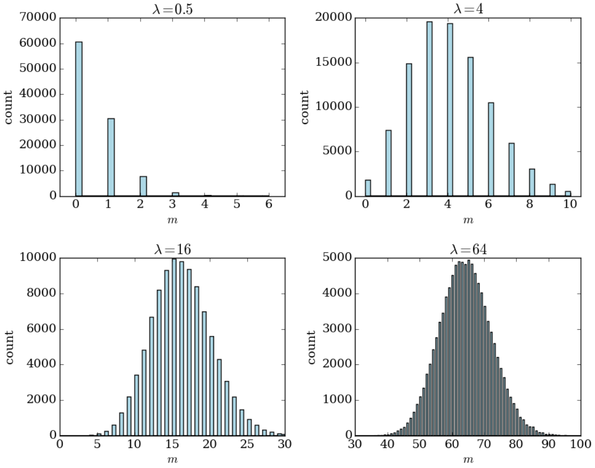
$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{n-m} \to \left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}$$

Combining the terms, we get the Poisson distribution:

$$p(m|\lambda) = \frac{e^{-\lambda}\lambda^m}{m!}$$

Poisson Distribution

The Poisson PDF is also discrete distribution:



Note how the Poisson distribution looks increasingly Gaussian as $\lambda \to$ large.

Poisson Distribution

Mean

The mean of the Poisson distribution is

$$\begin{aligned}\langle m \rangle &= \sum_{m=0}^{\infty} m \frac{\lambda^m e^{-\lambda}}{m!} \\ &= \lambda e^{-\lambda} \sum_{m=1}^{\infty} \frac{\lambda^{m-1}}{(m-1)!} \\ &= \lambda e^{-\lambda} \sum_{m'=0}^{\infty} \frac{m' \lambda^{m'}}{m'!} = \lambda e^{-\lambda} e^{\lambda} \\ &= \lambda\end{aligned}$$

where we used the fact that the sum is the expansion of e^{λ}.

Poisson Distribution

Variance

To find the variance $\mathrm{V}(m)$, we start with

$$\langle m(m-1)\rangle = \sum_{m=0}^{\infty} m(m-1)\cdot\frac{\lambda^m e^{-\lambda}}{m!}$$

As with the binomial distribution, drop the first two terms and set $m' = m - 2$ to get

$$\langle m^2 - m\rangle = \lambda^2 e^{-\lambda}\sum_{m'=0}^{\infty}\frac{\lambda^{m'}}{m'!} = \lambda^2$$

Therefore, the variance is

$$\begin{aligned}\mathrm{V}(m) &= \langle m^2\rangle - \langle m\rangle^2 = \langle m^2 - m\rangle + \langle m\rangle - \langle m\rangle^2 \\ &= \lambda^2 + \lambda - \lambda^2 \\ &= \lambda\end{aligned}$$

Poisson Distribution

HEP Example

Example

Suppose you try to measure a cross-section σ for a process.

- You observe n events for an integrated luminosity of $\mathcal{L}$.
- For this luminosity, the expected number of events is $\nu = \sigma\mathcal{L}$.
- The observed number of events will be Poisson-distributed according to ν.

Our best estimate of ν is the number of observed events: $\hat{\nu} = n$. For a Poisson distribution, the variance is equal to the mean, so uncertainty on our estimate is given by

$$\hat{\nu} = n \pm \sqrt{n} \quad \Longrightarrow \quad \hat{\sigma} = \hat{\nu}/\mathcal{L} = (n \pm \sqrt{n})/\mathcal{L}$$

Note: $\sqrt{n}$ is the *estimated* uncertainty of the underlying Poisson mean, not the uncertainty on n. There is no "error" on n, unless you miscounted!

Poisson Distribution

Neutrino Counts in Short Time Intervals

Example

From Barlow [1]: the number of neutrinos detected in 10-second intervals by the IMB detector on 23 February 1987 was:

No. events	0	1	2	3	4	5	6	7	8	9
No. intervals	1042	860	307	78	15	3	0	0	0	**1**

The prediction comes from a Poisson distribution with λ obtained by calculating the weighted average

$$\bar{m} = \hat{\lambda} = \sum_{i=0}^{8} w_i c_i / \sum_{i=0}^{8} w_i = \frac{0 \cdot 1042 + 1 \cdot 860 + \ldots}{1042 + 860 + \ldots} = 0.77$$

Given this mean, the expected Poisson counts are given by

Prediction	1064	823	318	82	16	2	0.3	0.03	0.003	0.0003

Combining Poisson Variables

Sum

The sum of two independent Poisson-distributed variables x and y is itself a Poisson variable z. To see this, first consider the joint probability of x and y:

$$p(x, y|\lambda_x, \lambda_y) = p(x|\lambda_x)p(y|\lambda_y) = \frac{e^{-\lambda_x}\lambda_x^x}{x!}\frac{e^{-\lambda_y}\lambda_y^y}{y!} = \frac{e^{-(\lambda_x+\lambda_y)}\lambda_x^x\lambda_y^y}{x!y!}$$

Now, to find $p(z|\lambda_z)$, sum $p(x, y)$ over all (x, y) satisfying $x + y = z$:

$$\begin{aligned}
p(z|\lambda_z) &= \sum_{x=0}^{z} \frac{e^{-(\lambda_x+\lambda_y)}\lambda_x^x\lambda_y^{z-x}}{x!(z-x)!} \\
&= \frac{e^{-(\lambda_x+\lambda_y)}}{z!}\sum_{x=0}^{z} \frac{z!\lambda_x^x\lambda_y^{z-x}}{x!(z-x)!} \\
&= \frac{e^{-(\lambda_x+\lambda_y)}}{z!}(\lambda_x+\lambda_y)^z, \quad \text{by the binomial theorem}
\end{aligned}$$

Combining Other Variables

Rules of the road:

- The sum of two Poisson variables is also a Poisson variable, even if the means are different.
- The sum of two Gaussian variables is a Gaussian, even if the means and variances are different.

This is not true for the binomial distribution. In this case:

$$\text{mean} = np_1 + Np_2, \qquad \text{variance} = np_1(1-p_1) + Np_2(1-p_2)$$

This does not have the general form of the binomial distribution unless $p_1 = p_2$. Also note:

- The difference of two Poissons is not Poisson; it follows a Skellam distribution.
- Beware of other false assumptions. E.g., the ratio of two Gaussians is not another Gaussian!

Gaussian Distribution

- You are already familiar with the Gaussian PDF:

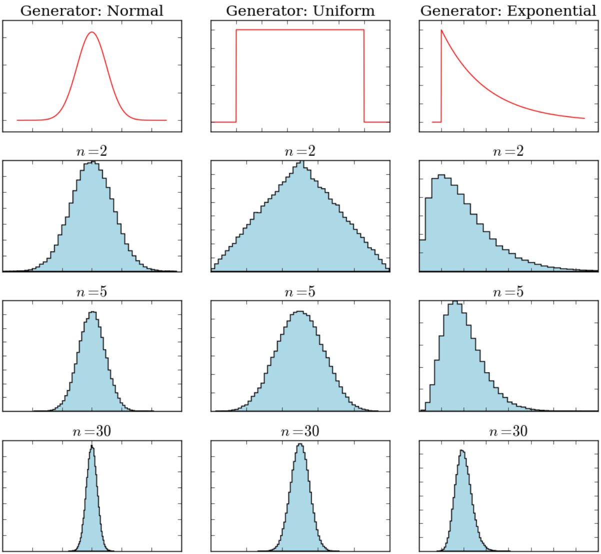
$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- The Gaussian is the limiting case of the Poisson distribution ($\lambda \to \infty$) and the binomial distribution ($n \to \infty$).
- Rules of thumb:
  - Poisson is a good approximation of binomial if $n \geq 20$ and $p \leq 0.05$.
  - Gaussian is a good approximation of Poisson if $\lambda \geq 20$.
  - Gaussian is a good approximation of binomial if $np(1-p) > 9$.
- So basically the Gaussian is usually "safe" for large numbers, but beware of using it in the wrong situation.
- The Gaussian has smaller tails than many other distributions and misusing it can cause you to overestimate the significance of rare events.

Central Limit Theorem

- Why is the Gaussian so important? Because of the Central Limit Theorem.
- **Theorem**: the sum of n independent continous random variables x_i with means μ_i and variances σ_i^2 becomes a Gaussian with mean and variance
$$\mu = \sum_{i=1}^{n} \mu_i \qquad \sigma^2 = \sum_{i=1}^{n} \sigma_i^2$$
in the limit $n \to \infty$.
- See Cowan [2] for a proof based on characteristic functions
- Generally, this is true independent of the individual forms of the PDFs of the x_i (see next slide).
- Since it is common for many measurements to add together in experiment, the Central Limit Theorem justifies the use of the Gaussian in many cases.

Central Limit Theorem



Multidimensional Gaussian

- The k-dimensional generalization of the Gaussian is

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- In this expression, $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ is a vector with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)$.
- $\boldsymbol{\Sigma}$ is the covariance matrix of the Gaussian. Its diagonal elements are the variances of the x_i, and its off-diagonal elements are the covariances $\text{cov}(x_i, x_j)$.

Example

Binormal distribution: for $k = 2$, $\boldsymbol{\Sigma}$ is a 2×2 real symmetric matrix:

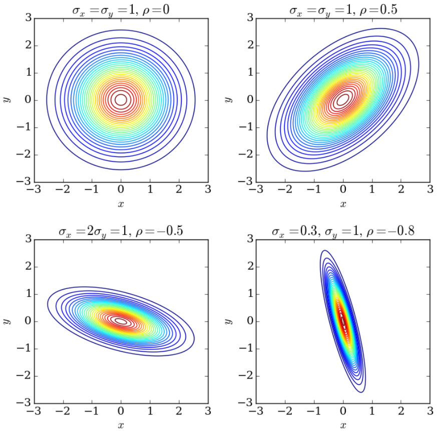
$$\boldsymbol{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

Change of Variables

- The covariance matrix fully specifies any correlations or anti-correlations between the elements of $\boldsymbol{x}$.
- If all of the elements of $\boldsymbol{x}$ are independent, then the covariance matrix is diagonal.
- If correlations exist, then there is a unitary matrix $\boldsymbol{U}$ that we can identify to diagonalize $\boldsymbol{\Sigma}$. I.e.,

$$\boldsymbol{\Sigma}' = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top.$$

- It is often convenient to change variables to $\boldsymbol{\Sigma}'$.



Uniform Distribution

- The uniform (a.k.a. the "top hat" distribution) has a probability which is constant inside some range $[a, b]$ and zero outside:

$$p(x|a,b) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \text{else} \end{cases}$$

- Mean: $\langle x \rangle = (a+b)/2$
- Variance: $\mathrm{V}(x) = (b-a)^2/12$
- Standard deviation: $\sigma_x = (b-a)/\sqrt{12}$
- The uniform distribution is important for two reasons:
  1. It is the basis for a large number of pseudorandom number generators.
  2. Its constant probability indicates no preferred values inside the range $[a, b]$, making it a popular "objective" prior probability density in Bayesian calculations.

χ^2 Distribution

- The χ^2 distribution of the continuous variable z is

$$p(z|n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2},$$

where Γ is the gamma function:

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$$

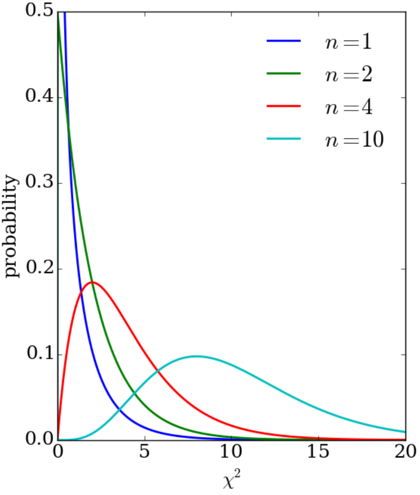
- Note: $\Gamma(x+1) = x\Gamma(x)$, and $\Gamma(1/2) = \sqrt{\pi}$. For integer x, $\Gamma(x+1) = x!$.
- **Mean**: $\mathrm{E}(x) = n$
- **Variance**: $\mathrm{V}(x) = 2n$
- The simple variance and mean of the χ^2 distribution make its tail probabilities easy to estimate.

χ^2 Distribution

- For n independent Gaussian x_i with means μ_i and variances σ_i^2, the quantity

$$z = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

  follows a χ^2 with n degrees of freedom.
- Notice that z looks like a least-squares estimator for a fit.
- Physicists often use the tail probability of χ^2 as a measure of goodness of fit.



Using the χ^2 Distribution

Example from S. Oser, UBC

Example

You are shown a fit and told that χ^2 is 70 for 50 degrees of freedom. Is the fit any good? In other words, how likely is it that χ^2 could be this large by chance?

Using the χ^2 Distribution

Example from S. Oser, UBC

Example

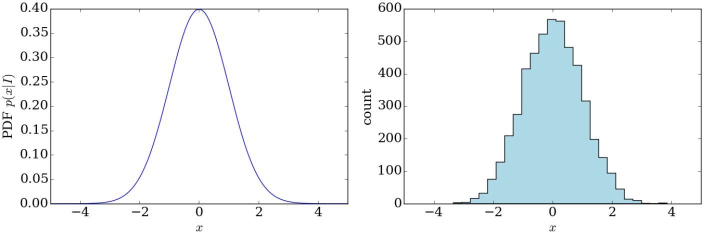
You are shown a fit and told that χ^2 is 70 for 50 degrees of freedom. Is the fit any good? In other words, how likely is it that χ^2 could be this large by chance?

Roughly: we expect the mean to be $n = 50$, and the variance is $2n = 100$ with RMS $\sqrt{100} = 10$. So this is a 2σ effect, which happens $\sim 2.5\%$ of the time if we approximate using the Gaussian definition of σ.

- If $\chi^2 \gg n$, then either your model is not a good fit to the data or you badly understimated your uncertainties σ_i.
- If $\chi^2 \ll n$, you should also be suspicious. You might have overestimated your uncertainties.

A Warning about Using the χ^2 Distribution

- Warning: the χ^2 statistic z is only asymptotically distributed like a χ^2 distribution if the uncertainties on each x_i are Gaussian.
- Where this can hurt you: fitting binned data.



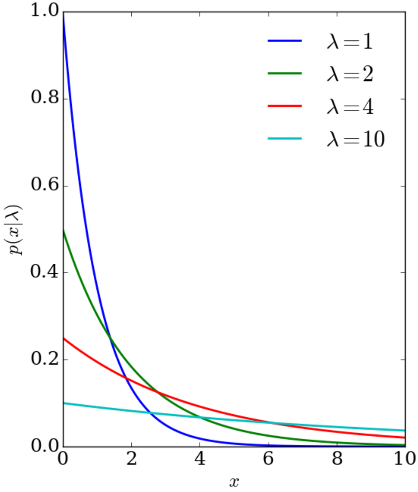
- Remember that if your histogram bins are relatively full the uncertainties on the counts in each bin will be Gaussian
- But if the bins are empty or close to empty, the uncertainties in the counts will be Poisson, and z will not follow the χ^2 distribution!

Exponential Distribution

- The exponential PDF is

$$p(x|\lambda) = \frac{1}{\lambda} e^{-x/\lambda}, \quad x \geq 0$$

- **Mean**: $E(x) = \lambda$.
- **Variance**: $V(x) = \lambda^2$, RMS: λ
- Lack of memory: $p(t - t_0 | t \geq t_0, \lambda) = p(t|\lambda)$.
- Decay time of unstable particle with lifetime $\lambda \to \tau$
- Lifetime of electrical components, such as lightbulbs



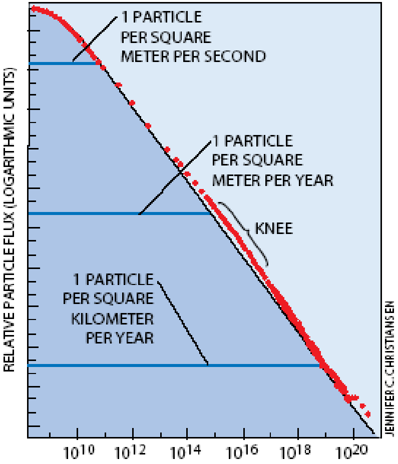
Power Law (Pareto) Distribution

- Power law:

$$p(x|\alpha) = Cx^{-\alpha}$$

- The power law shows up all over physics, and is characteristic of scale invariance, hierarchy, or stochastic generating processes.
- Examples: populations of cities, sizes of lunar impact craters, energies of cosmic rays, sizes of interstellar dust particles, magnitudes of earthquakes, ...

Scientific American, (c) 1998



Further Reading I

[1] R.J. Barlow. *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*. New York: Wiley, 1989.

[2] Glen Cowan. *Statistical Data Analysis*. New York: Oxford University Press, 1998.