Distributions derived from the normal distribution

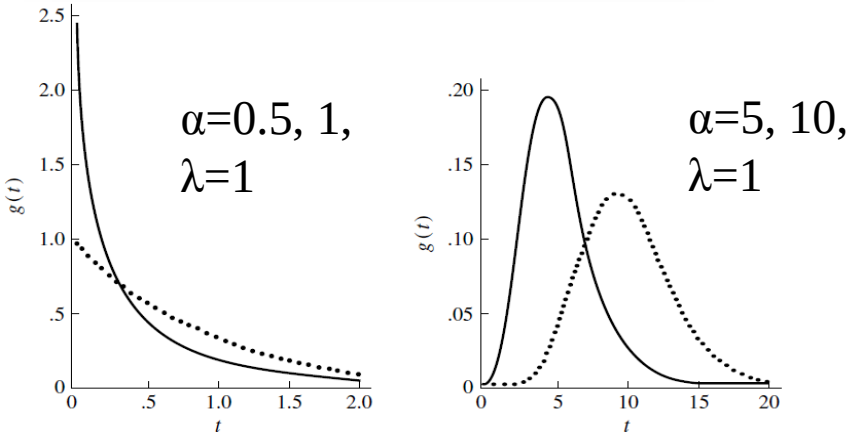
Chi-square χ^2, *t* and *F* distributions

DEFINITION

Our old friend!

If Z is a standard normal random variable, the distribution of $U = Z^2$ is called the chi-square distribution with 1 degree of freedom. ■

$$g(t) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}, \qquad t \geq 0$$



- Gamma distribution, α=λ=1/2, becomes χ_1^2 distribution with ***degree of freedom / d.o.f.*** = 1.

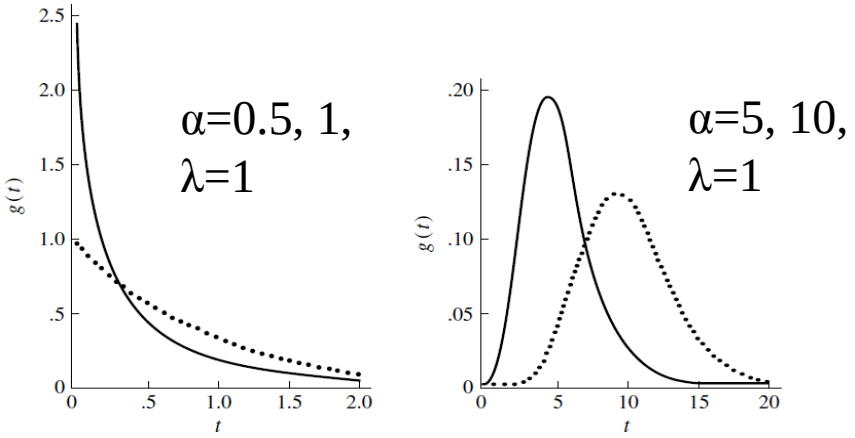
Chi-square χ^2, t and F distributions

DEFINITION

Our old friend!

If Z is a standard normal random variable, the distribution of $U = Z^2$ is called the chi-square distribution with 1 degree of freedom. ■

$$g(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}, \qquad t \geq 0$$



$$\Gamma\left(-\tfrac{3}{2}\right) = \tfrac{4}{3}\sqrt{\pi} \approx +2.363\,271\,801\,207$$
$$\Gamma\left(-\tfrac{1}{2}\right) = -2\sqrt{\pi} \approx -3.544\,907\,701\,811$$
$$\Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi} \approx +1.772\,453\,850\,906$$
$$\Gamma(1) = 0! = +1$$
$$\Gamma\left(\tfrac{3}{2}\right) = \tfrac{1}{2}\sqrt{\pi} \approx +0.886\,226\,925\,453$$
$$\Gamma(2) = 1! = +1$$
$$\Gamma\left(\tfrac{5}{2}\right) = \tfrac{3}{4}\sqrt{\pi} \approx +1.329\,340\,388\,179$$
$$\Gamma(3) = 2! = +2$$
$$\Gamma\left(\tfrac{7}{2}\right) = \tfrac{15}{8}\sqrt{\pi} \approx +3.323\,350\,970\,448$$
$$\Gamma(4) = 3! = +6$$

- Gamma distribution, $\alpha=\lambda=1/2$, becomes χ_1^2 distribution with ***degree of freedom / d.o.f.*** = 1.
- If $X \sim N(\mu, \sigma^2)$, then $(X-\mu)/\sigma \sim N(0, 1)$, and $[(X-\mu)/\sigma]^2 \sim \chi_1^2$.

Chi-square χ^2, t and F distributions

DEFINITION

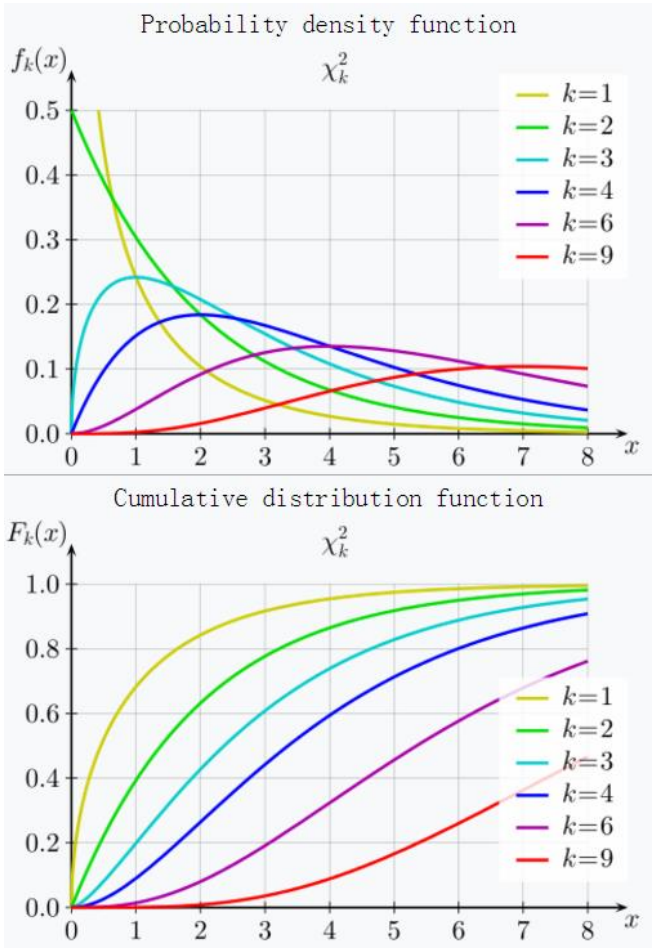
If $U_1, U_2, \ldots, U_n$ are independent chi-square random variables with 1 degree of freedom, the distribution of $V = U_1 + U_2 + \cdots + U_n$ is called the *chi-square distribution with n degrees of freedom* and is denoted by χ_n^2. ■

The sum of independent gamma random variables with same λ follows a gamma distribution. (Why?)

- Gamma distribution, $\alpha=n/2$, $\lambda=1/2$, becomes χ_n^2 distribution with d.o.f.=n.

$$f(v) = \frac{1}{2^{n/2}\Gamma(n/2)} v^{(n/2)-1} e^{-v/2}, \qquad v \geq 0$$

- Its mgf is $M(t)=(1-2t)^{-n/2}$
- Its **mean is *n*, variance is 2*n***
- **Reduced Chi-square** χ_n^2/n
- U and V are independent and $U \sim \chi_m^2$ and $V \sim \chi_n^2$, then $U+V \sim \chi_{m+n}^2$.



Chi-square χ^2, t and F distributions

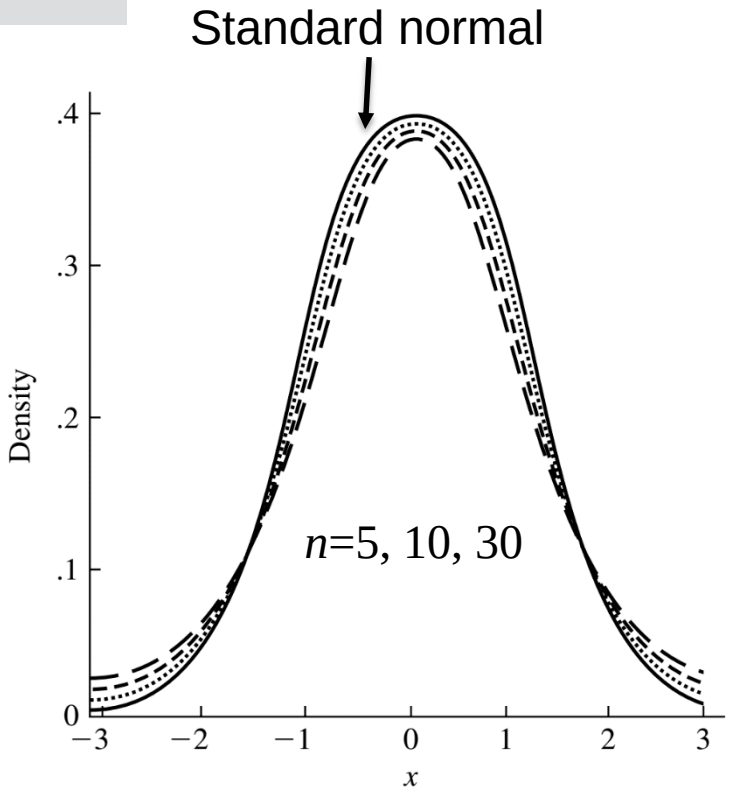
DEFINITION

If $Z \sim N(0, 1)$ and $U \sim \chi^2_n$ and Z and U are independent, then the distribution of $Z/\sqrt{U/n}$ is called the ***t* distribution** with n degrees of freedom. ■

$$f(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\,\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$$

- Note $f(t)=f(-t)$. Bell-shaped.
- As dof $\to \infty$, $t \to$ standard normal.
- For n>20-30, tails become lighter and the distributions are very close.

我们来讲点历史……



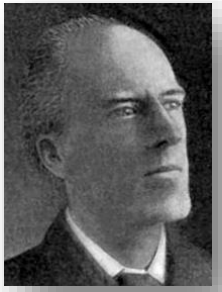
硝烟不断的20世纪统计学

Karl Pearson
(1857-1936)

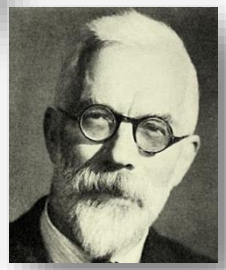
- 统计学带头大哥Karl Pearson (1857-1936)打击Bayes学派

硝烟不断的20世纪统计学

- 统计学带头大哥Karl Pearson (1857-1936)打击Bayes学派
- 推断统计学创始人Ronald Fisher (1890-1962，好斗)竟然排挤发展验证其理论的Jerzy Neyman, Egon Pearson（Karl之子）



Karl Pearson
(1857-1936)

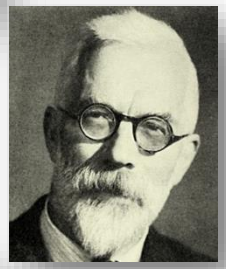


Ronald Fisher
(1890-1962)

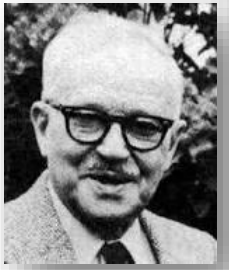
硝烟不断的20世纪统计学

- 统计学带头大哥Karl Pearson (1857-1936)打击Bayes学派
- 推断统计学创始人Ronald Fisher (1890-1962，好斗)竟然排挤发展验证其理论的Jerzy Neyman, Egon Pearson（Karl之子）
- 美国统计首领Jerzy Neyman与Fisher争斗，也打击Bayes学派

Karl Pearson
(1857-1936)



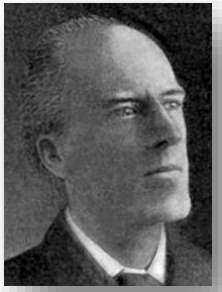
Ronald Fisher
(1890-1962)



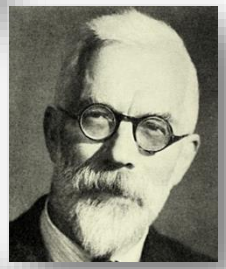
Jerzy Neyman
(1894-1981)

硝烟不断的20世纪统计学

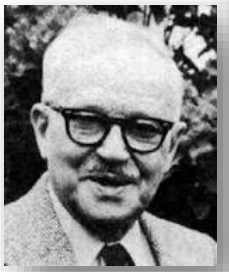
- 统计学带头大哥Karl Pearson (1857-1936)打击Bayes学派
- 推断统计学创始人Ronald Fisher (1890-1962，好斗)竟然排挤发展验证其理论的Jerzy Neyman, Egon Pearson（Karl之子）
- 美国统计首领Jerzy Neyman与Fisher争斗，也打击Bayes学派
- Bayes学派头号鼓吹者Leonard Savage也有强烈攻击性



Karl Pearson
(1857-1936)



Ronald Fisher
(1890-1962)

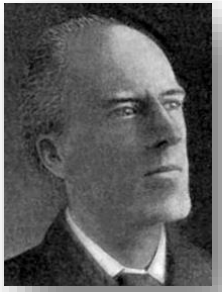


Jerzy Neyman
(1894-1981)

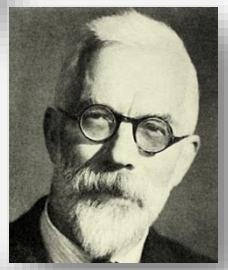
Leonard Savage
(1917-1971)

硝烟不断的20世纪统计学

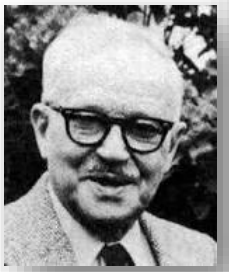
- 统计学带头大哥Karl Pearson (1857-1936)打击Bayes学派
- 推断统计学创始人Ronald Fisher (1890-1962，好斗)竟然排挤发展验证其理论的Jerzy Neyman, Egon Pearson（Karl之子）
- 美国统计首领Jerzy Neyman与Fisher争斗，也打击Bayes学派
- Bayes学派头号鼓吹者Leonard Savage也有强烈攻击性
- Gosset“居然”温和谦逊，关联Pearson和Fisher的工作，推动了推断统计学的诞生
- Guinness啤酒公司职员，公司不准发表成果。Pearson赏识他1904年的论文，急于要在其主编的《生物统计》发表。两人商量出student笔名，从此30多年，神秘的student陆续发表优秀论文，但很少人（只有活跃于英国的部分统计学者）知道他是谁。
- 1908年，Gosset以笔名student发表《平均数的规律误差》，为统计性推测的运用而发现了*t*分布。



Karl Pearson
(1857-1936)



Ronald Fisher
(1890-1962)



Jerzy Neyman
(1894-1981)

Leonard Savage
(1917-1971)

William Gosset
(1876-1937),
"**Student**"

Chi-square χ^2, t and F distributions

DEFINITION

Let U and V be independent chi-square random variables with m and n degrees of freedom, respectively. The distribution of

$$W = \frac{U/m}{V/n}$$

is called the F distribution with m and n degrees of freedom and is denoted by $F_{m,n}$. ■

$$f(w) = \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} w^{m/2-1} \left(1 + \frac{m}{n}w\right)^{-(m+n)/2}, \qquad w \geq 0$$

- For $n>2$, $E(W)$ exists, $=n/(n-2)$.

- **Make a guess!** What is the distribution of t^2?

t: Gaussian / $(\chi^2/n)^{1/2}$

Chi-square χ^2, t and F distributions

DEFINITION

Let U and V be independent chi-square random variables with m and n degrees of freedom, respectively. The distribution of

$$W = \frac{U/m}{V/n}$$

is called the F distribution with m and n degrees of freedom and is denoted by $F_{m,n}$. ■

$$f(w) = \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} w^{m/2-1} \left(1 + \frac{m}{n} w\right)^{-(m+n)/2}, \qquad w \geq 0$$

- For $n>2$, $E(W)$ exists, $=n/(n-2)$.

- **Make a guess!** What is the distribution of t^2?

t: Gaussian / $(\chi^2/n)^{1/2}$

t^2: Gaussian^2/(χ^2/n) or $(\chi^2/1)$ /$(\chi^2/n) \sim F_{1,n}$.

Limit Theorems,

the summit of probability theory

Under consideration here is the limiting behavior of the sum of independent random variables as the number of summands becomes large.

Many commonly computed statistical quantities, such as averages, can be represented as sums.

Tossing a coin many times successively: X_i takes on 0 or 1 according to whether the *i*th trial results in a tail or a head, and the proportion of heads in *n* trials is approaching ½,

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

THEOREM **A** *Law of Large Numbers*

Let $X_1, X_2, \ldots, X_i \ldots$ be a sequence of independent random variables with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$. Let $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i$. Then, for any $\varepsilon > 0$,

$$P(|\overline{X}_n - \mu| > \varepsilon) \to 0 \qquad \text{as } n \to \infty$$

Jacob Bernoulli's law of large numbers

THEOREM A *Law of Large Numbers*

Let $X_1, X_2, \ldots, X_i \ldots$ be a sequence of independent random variables with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$. Let $\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Then, for any $\varepsilon > 0$,

$$P(|\overline{X}_n - \mu| > \varepsilon) \to 0 \qquad \text{as } n \to \infty$$

样本够大，样本均值趋近于期望值。

Proof

We first find $E(\overline{X}_n)$ and $\text{Var}(\overline{X}_n)$:

$$E(\overline{X}_n) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \mu$$

Since the X_i are independent,

$$\text{Var}(\overline{X}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(X_i) = \frac{\sigma^2}{n}$$

The desired result now follows immediately from Chebyshev's inequality, which states that

$$P(|\overline{X}_n - \mu| > \varepsilon) \leq \frac{\text{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0, \qquad \text{as } n \to \infty \qquad \blacksquare$$

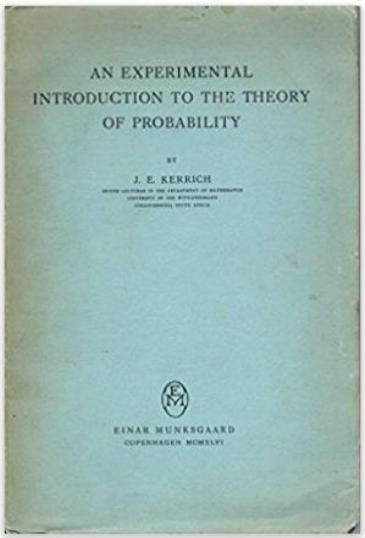
In the case of a fair coin toss, the X_i are Bernoulli random variables with $p = 1/2$, $E(X_i) = 1/2$ and $\mathrm{Var}(X_i) = 1/4$. If tossed 10,000 times

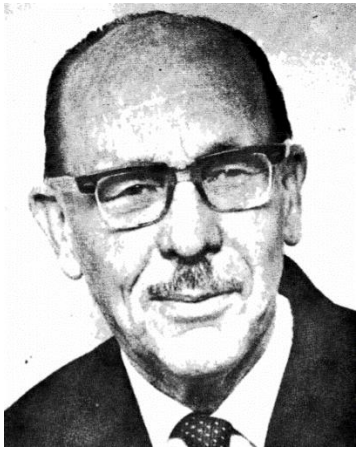
$$\mathrm{Var}(\overline{X}_{10,000}) = 2.5 \times 10^{-5}$$

Standard deviation is 0.005.

John Kerrich, a South African mathematician, tested this belief empirically while detained as a prisoner during World War II. He tossed a coin 10,000 times and observed 5067 heads.

Until the advent of computer simulations, his study, published in 1946, was widely cited as evidence of the asymptotic nature of probability. It is still regarded as a classic study in empirical mathematics.





John E. Kerrich (1903–1985) is noted for a series of experiments in probability he conducted while interned in Nazi-occupied Denmark in the 1940s.

About rigorousness…

If a sequence of random variables, $\{Z_n\}$, is such that $P(|Z_n - \alpha| > \varepsilon)$ approaches 0 as $n \to \infty$, for any $\varepsilon > 0$ and where α is some scalar, then Z_n is said to converge in probability to α.

The version of the law of large numbers stated and proved earlier asserts that X_n converges to μ in probability. This version is usually called the *weak law of large numbers.*

Under the same assumptions, a *strong law of large numbers*, which asserts that X_n converges almost surely to μ, can also be proved.

Strong convergence or *almost sure convergence:*
Z_n is said to converge almost surely to α if for every $\varepsilon > 0$, $|Z_n - \alpha| > \varepsilon$ only a finite number of times with probability 1; that is, beyond some point in the sequence, the difference is always less than ε, but where that point is random.

Monte Carlo Integration.

We wish to calculate

$$I(f) = \int_0^1 f(x)\,dx$$

Where f(x) is crazy. Generate independent uniform random variables on [0, 1], $X_1, X_2, \ldots, X_n$, compute

$$\hat{I}(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

When n is large, by the law of large numbers, this should be close to $E[f(X)]$,

$$E[f(X)] = \int_0^1 \mathbf{1}\cdot f(x)\,dx = I(f)$$

均匀分布

Compared to the standard numerical methods, not especially efficient in 1-d, but becomes increasingly efficient as the dimensionality grows.

Example:

$$I(f) = \frac{1}{\sqrt{2\pi}}\int_0^1 e^{-x^2/2}\,dx$$

If 1000 uniform points over [0, 1] are generated, one finds .3417 (exact value .3413).

$$\hat{I}(f) = \frac{1}{1000}\left(\frac{1}{\sqrt{2\pi}}\right)\sum_{i=1}^{1000} e^{-X_i^2/2}$$

Repeated measurements.

Unbiased measurements of a quantity, $X_1, X_2, \ldots, X_n$, are made. How close the average is to the true value μ depends not only n but on the variance of error, σ^2.

$$\boxed{\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}}$$

can be estimated: First,

$$n^{-1} \sum_{i=1}^{n} X_i^2 \rightarrow E(X^2)$$

Second, it can be shown that if Z_n, converges to α, g is a continuous function, then

$$g(Z_n) \rightarrow g(\alpha)$$

$$\overline{X}^2 \rightarrow [E(X)]^2$$

Finally,

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}^2 \rightarrow E(X^2) - [E(X)]^2 = \mathrm{Var}(X)$$

More generally, it follows from the law of large numbers that the sample moments converge in probability to the moments of X,

$$\boxed{n^{-1} \sum_{i=1}^{n} X_i^r \rightarrow E(X^r).}$$

生物学/医学实例：

肌肉或神经细胞膜有很多通道，这些通道打开时允许离子通过，关闭时不允许。

单个通道打开与否似乎是随机的。在平衡情形下，经常可假设通道打开与关闭相互独立，且仅有少数通道在任一时刻是打开的。

设通道打开的概率是*p*（很小），总共有*m*个通道（不知道），欲通过所有通道的总流量来测定单个通道可通过流量*c*。

在某时刻有N个通道打开，它是成功概率为p的m次试验所得二项随机变量。总流量为$S=cN$，可以直接测量。于是

$$E(S) = cE(N) = cmp$$

$$\mathrm{Var}(S) = c^2 mp(1-p)$$

$$\frac{\mathrm{Var}(S)}{E(S)} = c(1-p) \approx c$$

利用独立的测量值, $S_1, S_2, \ldots, S_n$, 可估算出其平均值和方差, 从而不需知道多少个通道的情况下，可估计出单个通道的流量c.

In applications, we often want to find $P(a<X<b)$ when we do not know the cdf of X precisely; it is sometimes possible to do this by approximating F_X. The approximation is often reached by some sort of limiting argument.

DEFINITION

Let $X_1, X_2, \ldots$ be a sequence of random variables with cumulative distribution functions $F_1, F_2, \ldots$, and let X be a random variable with distribution function F. We say that X_n converges in distribution to X if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

at every point at which F is continuous. ■

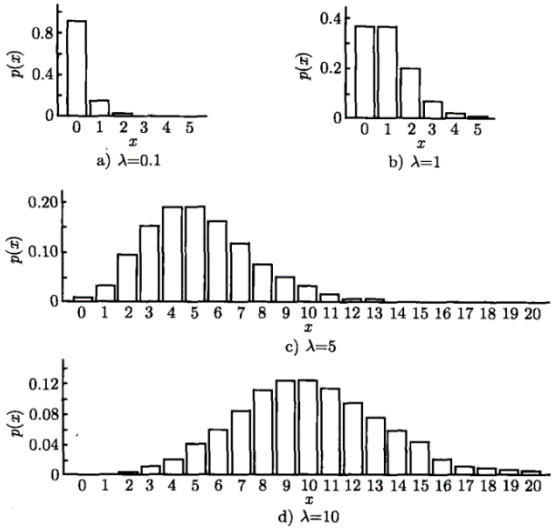
Mgfs are often useful for establishing the convergence of distribution functions. The unique determination between mgf and distribution holds for limits as well.

THEOREM A *Continuity Theorem*

Let F_n be a sequence of cumulative distribution functions with the corresponding moment-generating function M_n. Let F be a cumulative distribution function with the moment-generating function M. If $M_n(t) \to M(t)$ for all t in an open interval containing zero, then $F_n(x) \to F(x)$ at all continuity points of F. ■

Example.

Poisson can be approximated by Gaussian for large λ.



Let $\lambda_1, \lambda_2, \ldots$ be increasing with $\lambda_n \to \infty$, and let $\{X_n\}$ be a sequence of Poisson random variables with these parameters.

We know that $E(X_n) = \text{Var}(X_n) = \lambda_n$. The approximated Gaussian must have the same mean and variance as Poisson does. But, they are tending to infinity! Standardizing it,

$$Z_n = \frac{X_n - E(X_n)}{\sqrt{\text{Var}(X_n)}} = \frac{X_n - \lambda_n}{\sqrt{\lambda_n}}$$

Then $E(Z_n)=0$, $\text{Var}(Z_n)=1$. Its mgf should converge to standard normal.

$$M_{X_n}(t) = e^{\lambda_n(e^t - 1)}$$

$Y = a + bX$, then Y has the mgf $M_Y(t) = e^{at} M_X(bt)$

$$M_{Z_n}(t) = \mathrm{e}^{-t\sqrt{\lambda_n}} M_{X_n}\left(\frac{t}{\sqrt{\lambda_n}}\right) = \mathrm{e}^{-t\sqrt{\lambda_n}} \mathrm{e}^{\lambda_n(\mathrm{e}^{t/\sqrt{\lambda_n}} - 1)}$$

$$\log M_{Z_n}(t) = -t\sqrt{\lambda_n} + \lambda_n(e^{t/\sqrt{\lambda_n}} - 1)$$

$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$

$$\lim_{n\to\infty} \log M_{Z_n}(t) = \frac{t^2}{2} \qquad \lim_{n\to\infty} M_{Z_n}(t) = e^{t^2/2}$$

Example.
A certain type of particle is emitted at a rate of 900/hr. What is the probability that more than 950 particles will be emitted in a given hour if the counts form a Poisson process?

Let X be Poisson with mean 900. We find $P(X>950)$ by standardizing:

$$P(X > 950) = P\left(\frac{X - 900}{\sqrt{900}} > \frac{950 - 900}{\sqrt{900}}\right)$$

$$\approx 1 - \Phi\left(\tfrac{5}{3}\right)$$

$$= .04779$$

Exact value=0.4712.

A standardized Poisson random variable converges in distribution to a standard normal variable as λ approaches infinity.

Practically, we wish to use this limiting result as a basis for an approximation for large but finite values of λ.

For a good approximation, **λ does not have to be all that large (>10 or so).**

Central limit theorem (CLT)

Sum of random variables: $X_1, X_2, \ldots, X_n$ is a sequence of independent random variables with mean μ and variance σ^2.

$$S_n = \sum_{i=1}^{n} X_i$$

The law of large numbers tells us $S_n/n \to \mu$ in probability, since

$$\mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\mathrm{Var}(S_n) = \frac{\sigma^2}{n} \to 0$$

CLT is concerned not with how S_n/n fluctuates around μ. Standardizing it,

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

实为$(X-\mu)/\sigma$形式

Z_n has mean 0 and variance 1. CLT states that the distribution of it converges to the standard normal distribution.

Central limit theorem (CLT)

Let $X_1, X_2, \ldots$ be a sequence of independent random variables having mean 0 and variance σ^2 and the common distribution function F and moment-generating function M defined in a neighborhood of zero. Let

$$S_n = \sum_{i=1}^{n} X_i$$

Then

$$\lim_{n\to\infty} P\left(\frac{S_n}{\sigma\sqrt{n}} \le x\right) = \Phi(x), \qquad -\infty < x < \infty$$

Standard normal cdf

定理 5.3.2 (中心极限定理) 令 $X_1, X_2, \cdots$ 是均值为 0 和方差为 σ^2 的独立随机变量序列，具有相同的分布函数 F，矩生成函数 M 在零点附近有定义. 令

$$S_n = \sum_{i=1}^{n} X_i$$

那么

$$\lim_{n\to\infty} P\left(\frac{S_n}{\sigma\sqrt{n}} \leqslant x\right) = \Phi(x), \quad -\infty < x < \infty$$

Central limit theorem (CLT)

Proof. Let $Z_n = S_n/(\sigma\sqrt{n})$. We will show its mgf → standard normal.

$$S_n = \sum_{i=1}^{n} X_i \quad \rightarrow \quad M_{S_n}(t) = [M(t)]^n \quad \rightarrow \quad M_{Z_n}(t) = \left[M\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$$

Expand $M(s)$ about 0,

$$M(s) = M(0) + sM'(0) + \frac{1}{2}s^2M''(0) + \varepsilon_s$$

$E(X) = 0,\ M'(0) = 0,\ M''(0) = \sigma^2$

When $n \to \infty$, $t/(\sigma\sqrt{n}) \to 0$, **Why?**

$$M\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 + \frac{1}{2}\sigma^2\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + \varepsilon_n$$

$$M_{Z_n}(t) = \left(1 + \frac{t^2}{2n} + \varepsilon_n\right)^n$$

$$\lim_{n\to\infty}\left(1 + \frac{a_n}{n}\right)^n = \mathrm{e}^a$$

$$M_{Z_n}(t) \to \mathrm{e}^{t^2/2}$$

- Standard normal mgf.

Central limit theorems

There are many central limit theorems of various degrees of abstraction and generality. The above is one of the simplest version of CLT.

Relaxing various assumptions:

1. It would only be necessary that 1st and 2nd moments exist. (The existence of mgf is a strong assumption, one can use characteristic functions).
2. Further generalizations weaken the assumption that X_i have the same distribution, and apply to linear combinations of independent variables.
3. CLTs have also been proved that weaken the independence assumption and allow X_i to be dependent to some extent.

Central limit theorems

There are many central limit theorems of various degrees of abstraction and generality. The above is one of the simplest version of CLT.

Relaxing various assumptions:

1. It would only be necessary that 1st and 2nd moments exist. (The existence of mgf is a strong assumption, one can use characteristic functions).
2. Further generalizations weaken the assumption that X_i have the same distribution, and apply to linear combinations of independent variables.
3. CLTs have also been proved that weaken the independence assumption and allow X_i to be dependent to some extent.

For practical purposes, itself is not of primary interest. Statisticians are more interested in its use as an approximation with finite values of *n*.

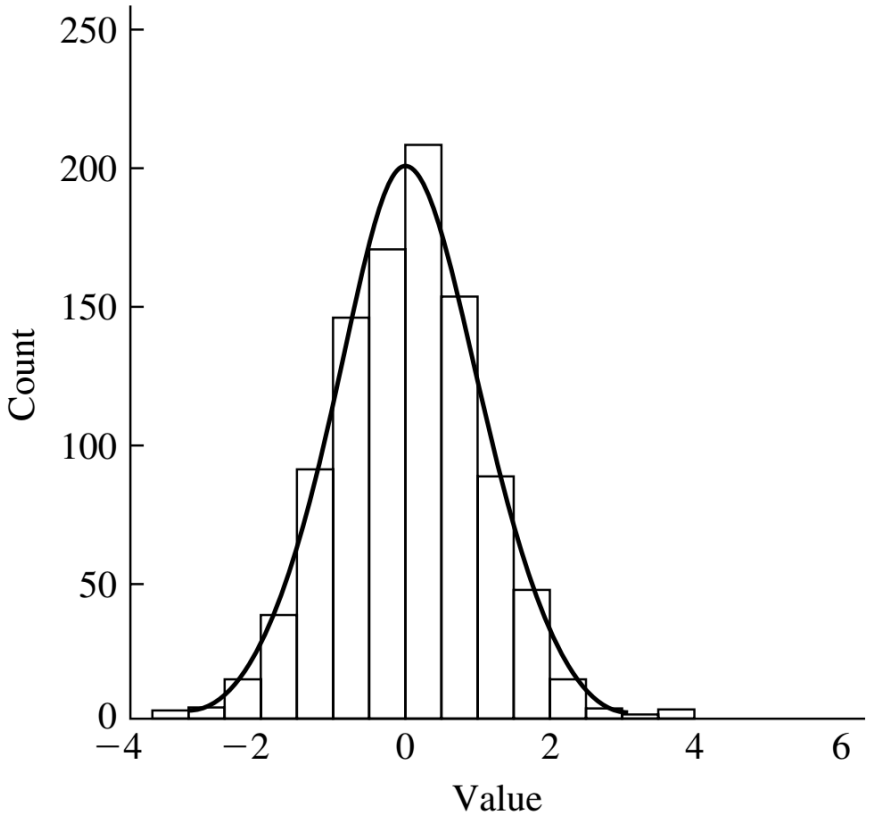
How fast the approximation becomes good depends on the distribution.

- *If it is fairly symmetric and has tails that die off rapidly, the approximation becomes good for relatively small values of n.*
- *If the distribution is very skewed or if the tails die down very slowly, a larger value of n is needed for a good approximation.*

Example: approximation

Because uniform distribution on [0, 1] has mean 1/2 and variance 1/12 , the sum of 12 uniform random variables, minus 6, has mean 0 and variance 1.

The distribution of this sum is quite close to normal; in fact, *before better algorithms were developed, it was commonly used in computers for generating normal random variables from uniform ones.*

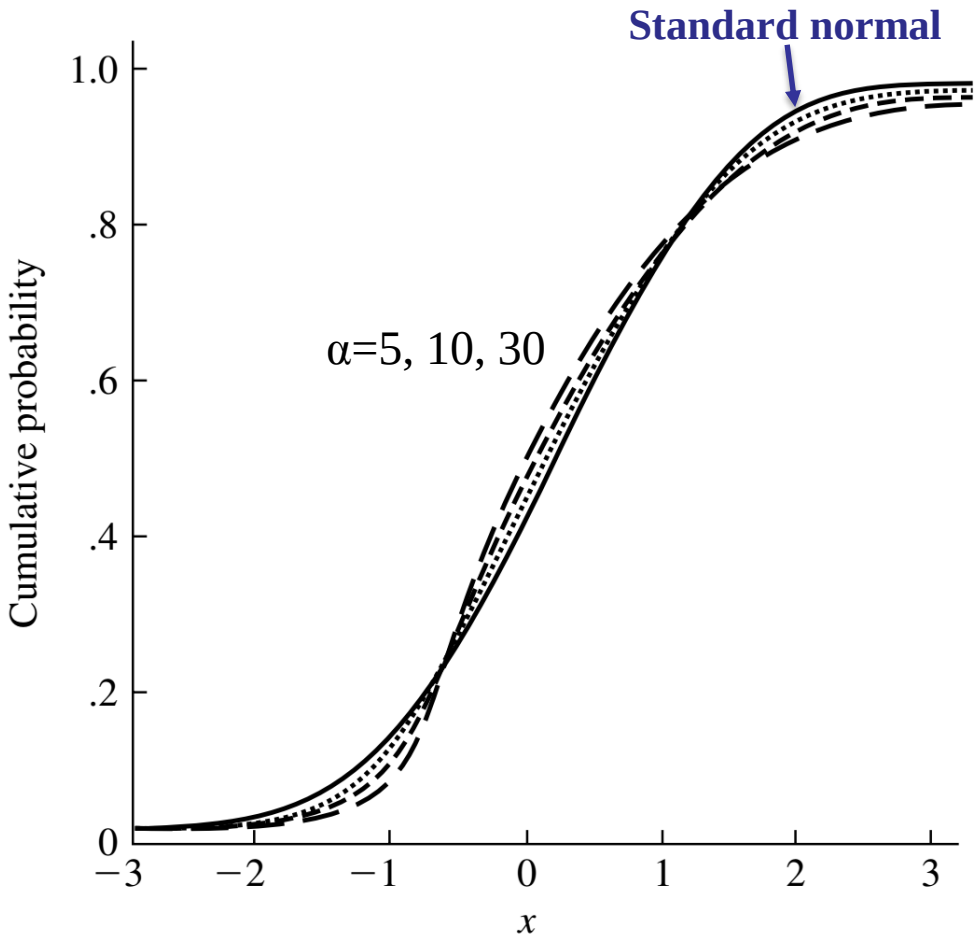


A histogram of 1000 values, each of which is the sum of 12 uniform [-1/2,1/2] pseudorandom variables, with an approximating standard normal density.

The fit is surprisingly good, despite that 12 is *not that large* a value of *n*.

Example: approximation

The sum of n independent exponential random variables with $\lambda = 1$ follows a gamma distribution with $\lambda = 1$ and $\alpha = n$. The exponential density is quite skewed; therefore, a good approximation of a standardized gamma by a standardized normal would not be expected for small n.



The cdf's of the standard normal and standardized gamma distributions for increasing values of n.

The approximation improves as n increases.

Example: measurement error.

$X_1, X_2, \ldots, X_n$ are repeated, independent measurements of a quantity, μ, and that $E(X_i)=\mu$ and $\mathrm{Var}(X_i)=\sigma^2$. The average, $\overline{X}$, is an estimate of μ. Law of large numbers makes us hope that the average is close to μ for large n.

Chebyshev's inequality allows for bounding the probability of an error of a given size, CLT gives a much sharper approximation to the actual error.

Suppose we wish to find $P(|\overline{X} - \mu| < c)$ for some constant c. To use CLT, we standardize, using $E(\overline{X}) = \mu$ and $\mathrm{Var}(\overline{X}) = \sigma^2/n$:

$$P(|\overline{X} - \mu| < c) = P(-c < \overline{X} - \mu < c)$$

$$= P\left(\frac{-c}{\sigma/\sqrt{n}} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \frac{c}{\sigma/\sqrt{n}}\right)$$

$$\approx \Phi\left(\frac{c\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{c\sqrt{n}}{\sigma}\right)$$

Example: measurement error.

$X_1, X_2, \ldots, X_n$ are repeated, independent measurements of a quantity, μ, and that $E(X_i)=\mu$ and $\text{Var}(X_i)=\sigma^2$. The average, $\overline{X}$, is an estimate of μ. Law of large numbers makes us hope that the average is close to μ for large n.

Chebyshev's inequality allows for bounding the probability of an error of a given size, CLT gives a much sharper approximation to the actual error.

Suppose we wish to find $P(|\overline{X} - \mu| < c)$ for some constant c. To use CLT, we standardize, using $E(\overline{X}) = \mu$ and $\text{Var}(\overline{X}) = \sigma^2/n$:

$$P(|\overline{X} - \mu| < c) = P(-c < \overline{X} - \mu < c)$$

$$= P\left(\frac{-c}{\sigma/\sqrt{n}} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \frac{c}{\sigma/\sqrt{n}}\right)$$

$$\approx \Phi\left(\frac{c\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{c\sqrt{n}}{\sigma}\right)$$

Suppose that 16 measurements are taken with σ=1. The probability that the average deviates from μ by less than .5 is approximately

$$P(|\overline{X} - \mu| < .5) = \Phi(.5 \times 4) - \Phi(-.5 \times 4) = .954$$

The reasoning can be turned around: given c and γ, find n such that

$$P(|\overline{X} - \mu| < c) \geq \gamma$$

Example: Normal approximation to Binomial.

A Binomial random variable is the sum of independent Bernoulli random variables → its distribution can be approximated by Gaussian.

The approximation is best when p=?

Example: Normal approximation to Binomial.

A Binomial random variable is the sum of independent Bernoulli random variables → its distribution can be approximated by Gaussian.

The approximation is best when p=1/2, due to symmetry.

A frequently used ***rule of thumb***: approximation is reasonable when

$$\boldsymbol{n\,p > 5} \text{ and } \boldsymbol{n\,(1-p) > 5.}$$

Question: If a coin is tossed 100 times and lands heads up 60 times. Should we be surprised and doubt that the coin is fair?

Example: Normal approximation to Binomial.

A Binomial random variable is the sum of independent Bernoulli random variables → its distribution can be approximated by Gaussian.

The approximation is best when p=1/2, due to symmetry.

A frequently used ***rule of thumb***: approximation is reasonable when

$$\boldsymbol{np > 5} \text{ and } \boldsymbol{n(1-p) > 5.}$$

Question: If a coin is tossed 100 times and lands heads up 60 times. Should we be surprised and doubt that the coin is fair?

Method 1: If the coin is fair, # of heads, X, is a binomial random variable with n = 100 trials and p = 1/2, so that $E(X) = np = 50$ and $\text{Var}(X) = np(1-p) = 25$. We could calculate $P(X = 60)$, which would be a small number. But because there are so many possible outcomes, $P(X = 50)$ is also a small number, so this calculation would not really answer the question.

Example: Normal approximation to Binomial.

A Binomial random variable is the sum of independent Bernoulli random variables → its distribution can be approximated by Gaussian.

The approximation is best when p=1/2, due to symmetry.

A frequently used ***rule of thumb***: approximation is reasonable when

$$\boldsymbol{np > 5} \text{ and } \boldsymbol{n(1-p) > 5.}$$

Question: If a coin is tossed 100 times and lands heads up 60 times. Should we be surprised and doubt that the coin is fair?

Method 1: If the coin is fair, # of heads, X, is a binomial random variable with n = 100 trials and p = 1/2, so that $E(X) = np = 50$ and $\text{Var}(X) = np(1-p) = 25$. We could calculate $P(X = 60)$, which would be a small number. But because there are so many possible outcomes, $P(X = 50)$ is also a small number, so this calculation would not really answer the question.

Method 2: Instead, we calculate the probability of a deviation as extreme as or more extreme than 60 if the coin is fair; that is, we calculate $P(X \geq 60)$. To approximate this probability from the normal distribution, we standardize

$$\begin{aligned} P(X \geq 60) &= P\left(\frac{X-50}{5} \geq \frac{60-50}{5}\right) \\ &\approx 1 - \Phi(2) \\ &= .0228 \end{aligned}$$

Example: Particle size distribution.

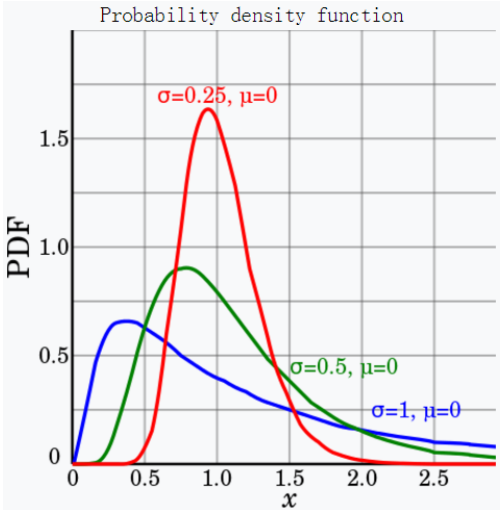
The size distribution of grains of particulate matter is often quite skewed, with a slowly decreasing right tail. The **lognormal distribution** can be fit to it (meaning log X has a normal distribution). CLT gives a theoretical rationale for the use of the lognormal distribution in some situations.

Suppose that a particle of initial size y_0 is subjected to repeated impacts, that on each impact a proportion, X_i, of the particle remains, and that the X_i are modeled as independent random variables having the same distribution. After the first impact, the size of the particle is $Y_1 = X_1 y_0$; after the second impact, the size is $Y_2 = X_2 X_1 y_0$; and after the nth impact, the size is

$$Y_n = X_n X_{n-1} \cdots X_2 X_1 y_0$$

$$\log Y_n = \log y_0 + \sum_{i=1}^{n} \log X_i$$

CLT applies to $\log Y_n$.



$$\ln \mathcal{N}(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right]$$

Example: theory of finance.

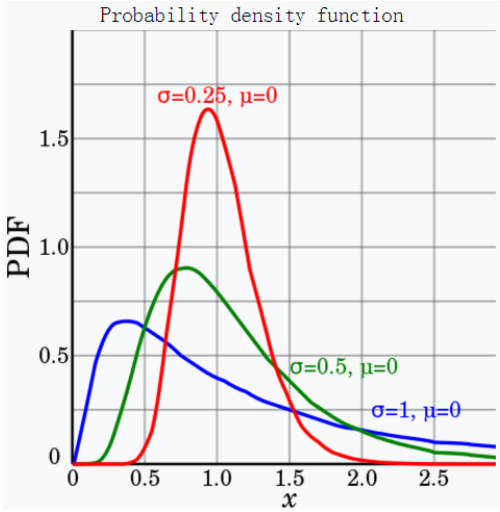
A similar construction is relevant to the theory of finance. An initial investment of value v_0 is made and returns occur in discrete time, i.e. daily.

If the return on the first day is R_1, then the value becomes $V_1 = R_1 v_0$. After day two the value is $V_2 = R_2 R_1 v_0$, and after day n the value is

$$V_n = R_n R_{n-1} \cdots R_1 v_0$$

$$\log V_n = \log v_0 + \sum_{i=1}^{n} \log R_i$$

CLT: If the returns are independent random variables with the same distribution, then the distribution of $\log V_n$ is approximately normally distributed.



$$\ln \mathcal{N}(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right]$$

"I have had my results for a long time: but I do not yet know how I am to arrive at them."

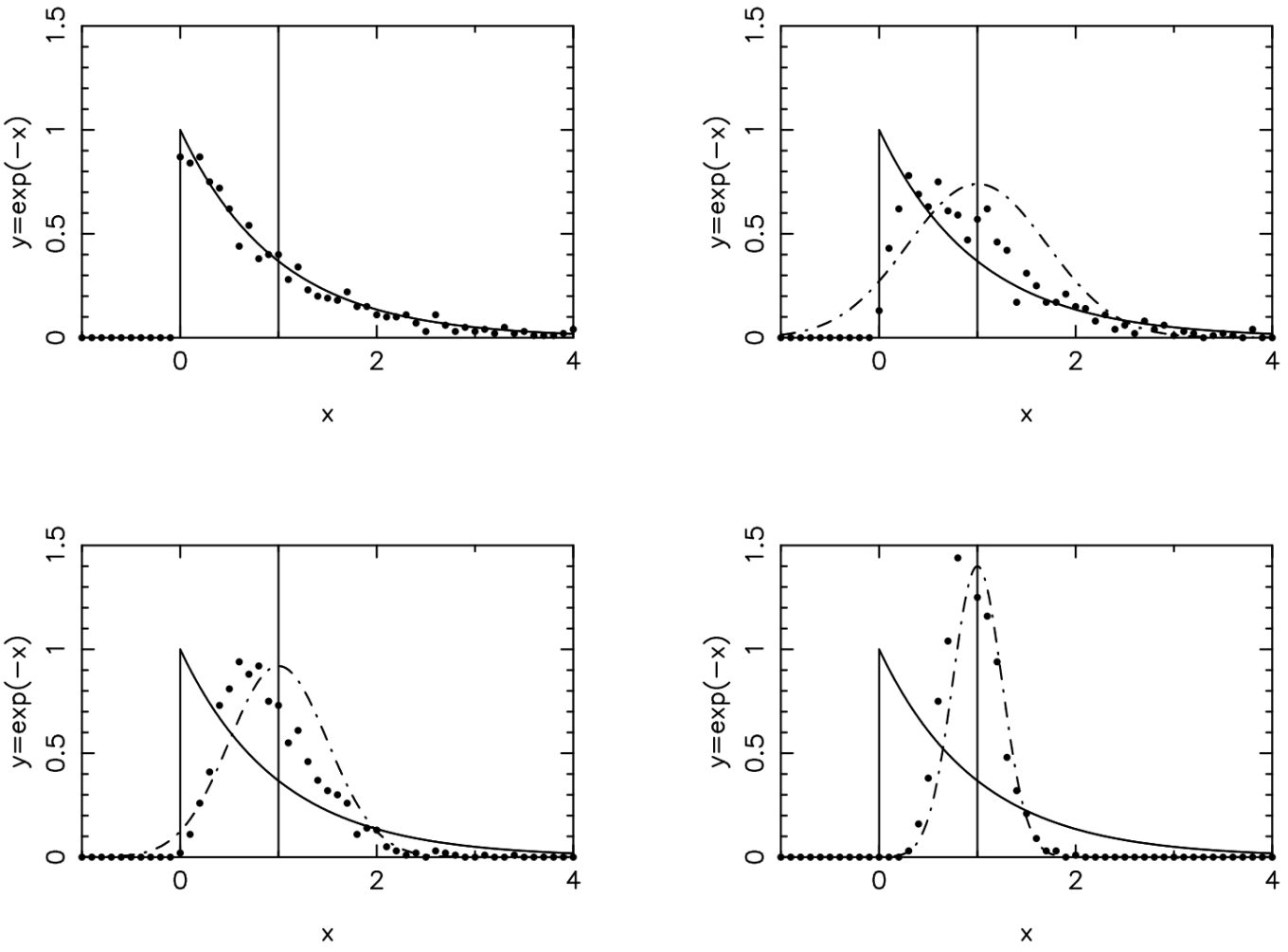
CLT is among the most remarkable theorems ever

• A little bit of summing/averaging will produce a Gaussian distribution of results **no matter the shape of distribution from which the sample is drawn**.

• Errors on averaged samples will always look "Gaussian".

CLT shapes our entire view of experimentation.
=> error language of sigmas, describing tails of Gaussian distributions

An indication of the compelling power of CLT. The panels show successive amounts of “integration”: a) a single value has been drawn; b) 200 values have been taken from an average of two values; c) 200 values from an average of four; d) 200 values from an average of 16.



Tasty appetizers:

1. a fishing trip

Correlation - why do we try it?

When we make a set of measurements, it is instinct to try to correlate the observations with other results. We might wish

(1) to check that other observers' measurements are reasonable,

(2) to check that our measurements are reasonable,

(3) to test a hypothesis, perhaps one for which the observations were explicitly made,

(4) in the absence of any hypothesis, any knowledge, or anything better to do with the data, to find if they are correlated with other results in the hope of discovering some New and Universal Truth.

We are gonna do it – and we are going to fall into some deadly traps. We already have.

The fishing trip

Suppose that we have plotted something against something, on a Fishing Expedition.

The fishing trip

Suppose that we have plotted something against something, on a Fishing Expedition.

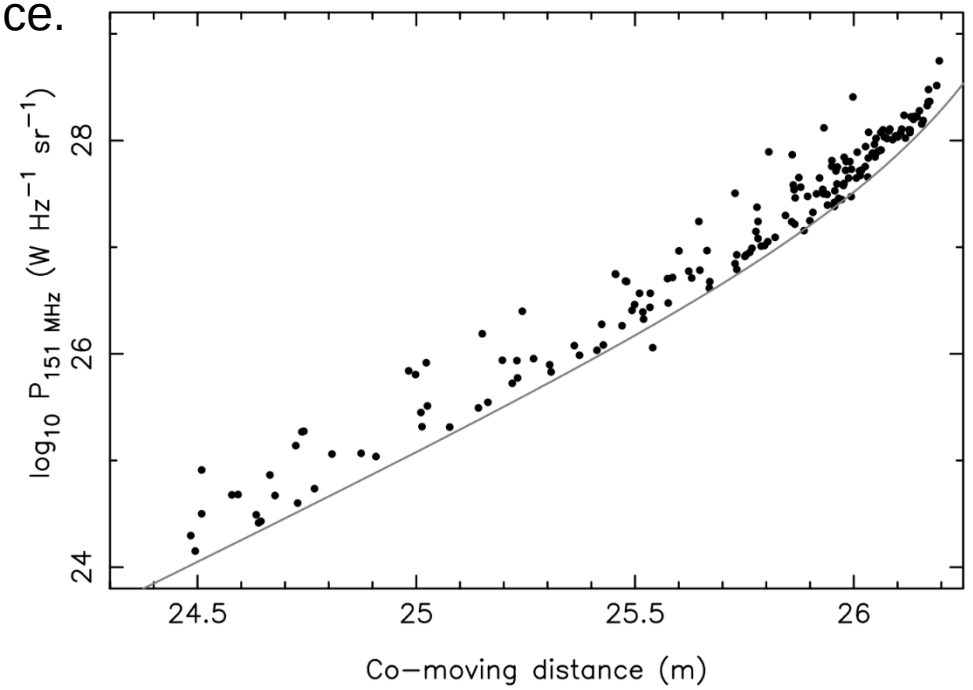
-Expedition是“远航、考察”。可能来自出海探寻鱼群行踪的远征。鱼儿潜游在茫茫大海里，捉摸不定，找到他们得靠机遇。引伸：以搜罗挖掘不利于某人的证据为目的的调查行动。The senator says the investigation is **a fishing expedition** by his enemies to see if they can find anything he has ever done that might hurt his political career.

The fishing trip

Suppose that we have plotted something against something, on a Fishing Expedition.

1. Does the eye see much correlation? If not, formal testing for correlation is probably a waste of time. *The eyeball is an excellent statistical device.*

2. Could the apparent correlation be due to selection effects? Consider for instance the beautiful correlation obtained by Sandage (1972): 3CR radio luminosities vs distance.



Radio luminosities of 3CR radio sources versus distance modulus

Still on the fishing trip …

The plot proves luminosity evolution for radio sources? Are the more distant objects (at earlier epochs) clearly not the more powerful?

No! The sample is flux- (or apparent intensity) limited; the solid line shows the flux-density limit of the 3CR catalog. The lower right-hand region can never be populated; such objects are too faint to show above the limit of the 3CR catalog.

But the upper left? Provided that *the luminosity function* (the true space density in objects per Mpc3) *slopes downward* with increasing luminosity, the objects are bound to crowd towards the line.

This is the *only conclusion* to be drawn from the diagram!

Still on the fishing trip …

Astronomers produce many plots of this type, and say things like "The lower right-hand region of the diagram is unpopulated because of the detection limit, but there is no reason why objects in the upper left-hand region should have escaped detection...."

Nonsense – probabilities rule! There are only low-luminosity sources to be seen at low redshifts because there's not enough volume to pick up the high-luminosity counterparts.

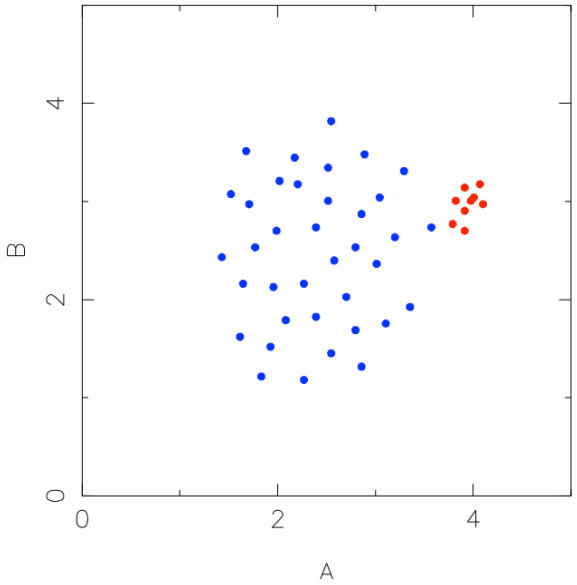
This applies to any proposed correlation for variables with steep probability functions dependent upon one of the variables plotted.

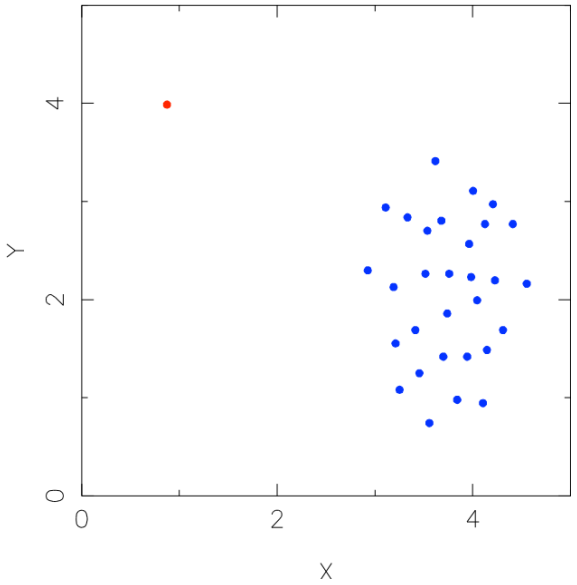
-- "***Malmquist bias***"

Still fishing …

3. If we are happy about (2), we can try formal calculation of the significance of the correlation. But, if there is a correlation, does the regression line (the fit) make sense?

4. If we are still happy - is the formal result realistic?
Rule of Thumb – *if* ***10% of the points*** *are grouped by themselves so that covering them with the thumb destroys the correlation to the eye, then we should doubt it.* Selection effects, data errors, or some other form of statistical conspiracy?





Suspect correlations: in each case formal calculation will indicate that a correlation exists to a high degree of significance!

Fishing, fishing …

5. If still confident, remember that
a correlation does not prove a causal connection. Examples:

- The price of fish in Walmart Market and the size of feet in China.
- Number of violent crimes in cities versus number of churches.
- The quality of student handwriting versus their height.
- Stock market prices and the sunspot cycle.
- Cigarette smoking versus lung cancer.
- Health versus alcohol intake…

1. Lurking third variables
2. Similar time scales
3. Causal connection…

There are ways of searching for intrinsic correlation between variables when they are known to depend mutually upon a third variable.

But… "known"???

Wilkinson & Pickett: *The Spirit Level*

"Correlations" show that higher income inequality correlates with higher crime rate, higher infant mortality, lower life expectancy, worse gender inequality, lower education standards, higher obesity (肥胖) rates……

Figure 5a: Wilkinson and Pickett's plot of inequality against homicide rates[33]

(谋杀)

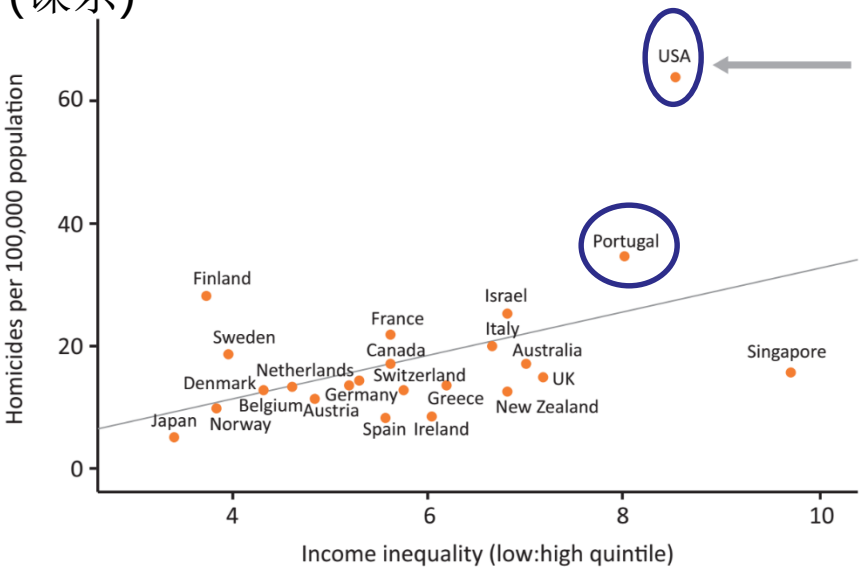
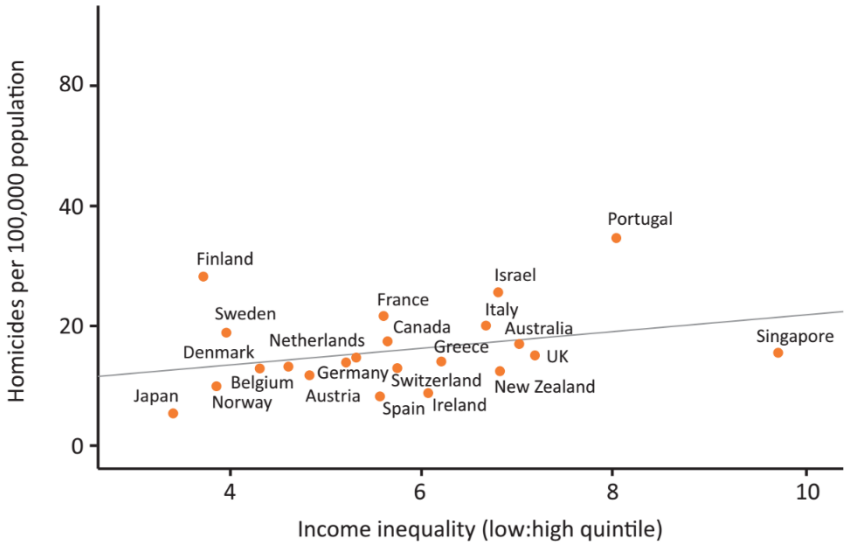


Figure 5b: Wilkinson and Pickett's plot of inequality against homicide rates, excluding the USA

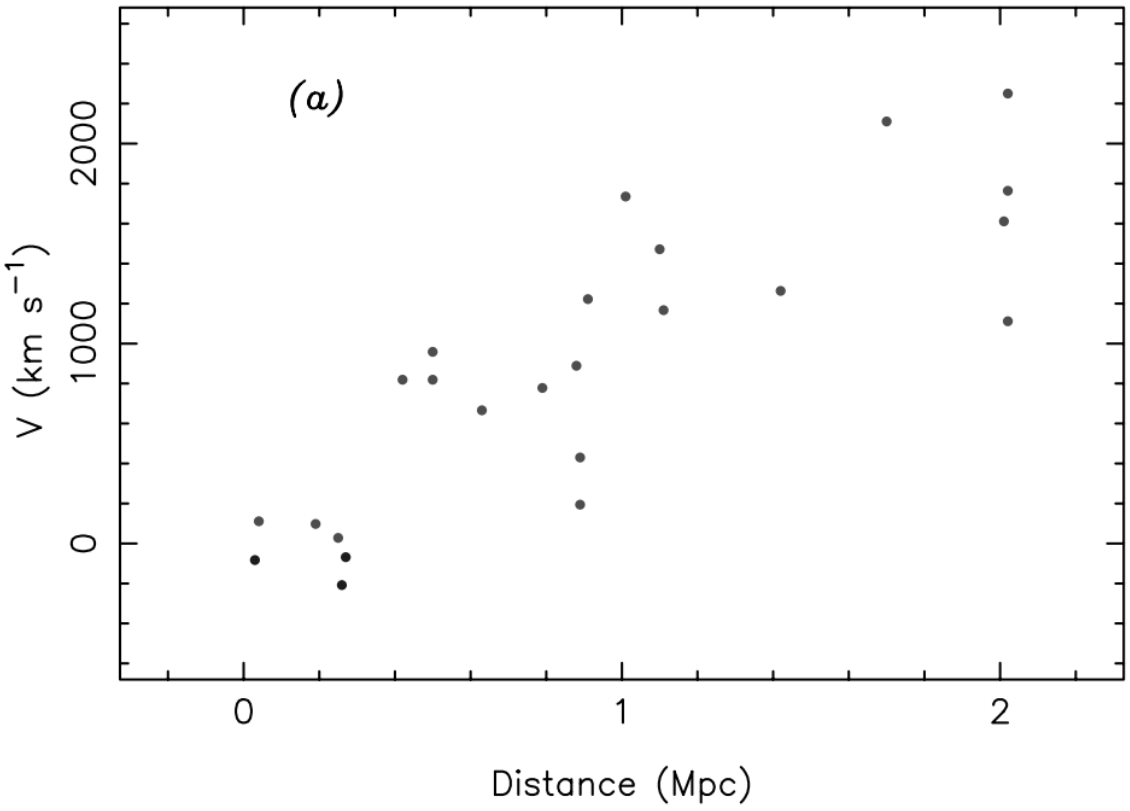


Critique by Peter Saunders: *Beware of False Prophets*（先知）shows that it is (statistical) garbage. The "correlations" are false or of no significance. The data are selective.

"Conclusion: There is no evidence of a significant association between the level of income inequality in a country and its homicide rate."

The end of the fishing trip – big fish are out there

Don’t get too discouraged by all the foregoing. Consider the example figure, a ragged correlation if ever there was one, although there are no nasty groupings of the type rejected by the Rule of Thumb.



An early Hubble diagram (Hubble 1936); recession velocities of a sample of 24 galaxies versus distance measure.

Formal correlation analysis later…

Tasty appetizers:

2. power law distributions

Integral form: $N(>L) = K\,L^{\gamma+1}$, or

Differential form: $dN = (\gamma+1)\,K\,L^{\gamma}\,dL$

Scale-free or ***scale-independent*** distribution:
If $f(x)=x^{\gamma}$, then $f(ax)= a^{\gamma}\cdot x^{\gamma} = \text{const}\cdot x^{\gamma} = \text{const}\cdot f(x)$

- Distribution of fluctuations in the economic market
- Growth rates of firms
- Distribution of salaries
- Size distribution of avalanches (雪崩), earthquakes and forest fires.

Criticality:
- The onset of avalanches
- Similarly, adding sand at the apex of a sand-pile to a point where it suddenly becomes unstable
- At criticality, no prescription as to whether a small region will slide and stop, or whether the entire side of the pile will be collectively triggered to break away and collapse…

Criticality:

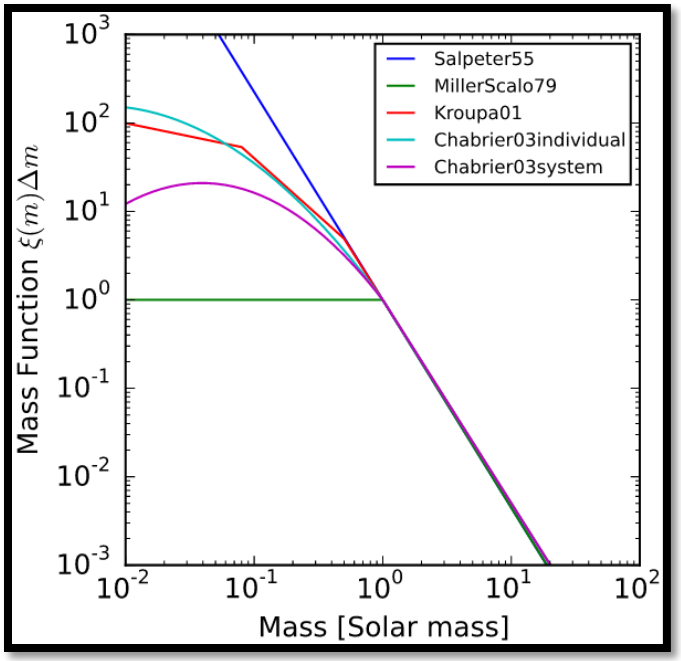
- earthquakes, stock-market fluctuations, forest fires, sub-networks on the internet (scale-free system).
- The main feature in each case is a *negative exponent*. There are many more small things than large, many little sandslips go nowhere while very rare are the catastrophic（灾难性）collapses of the pile.

- not formally a probability distribution because $\int = \infty$, infinite mean & variance!
- normally there are physical bounds so that it works
- Experience gained from ordinary pdfs and Gaussian now becomes misleading…

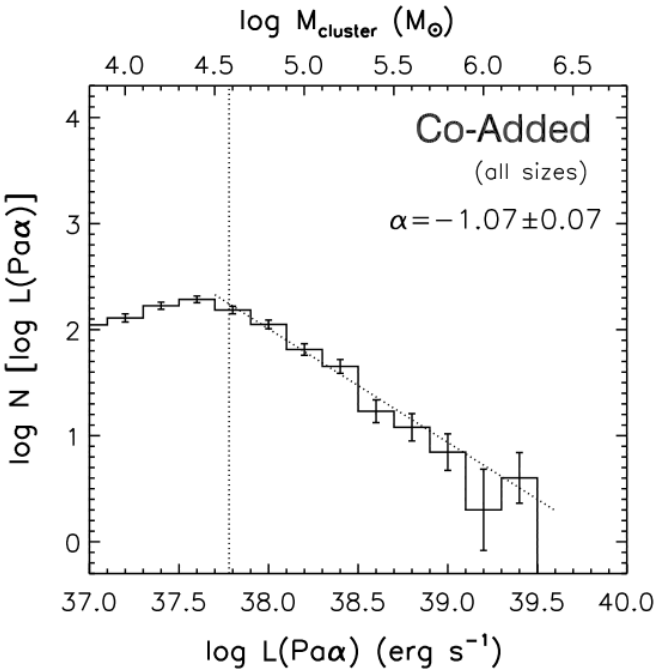
In astronomy, slope is invariably negative, steep power laws $-4 < \gamma < 0$ pop up in astronomy frequently.

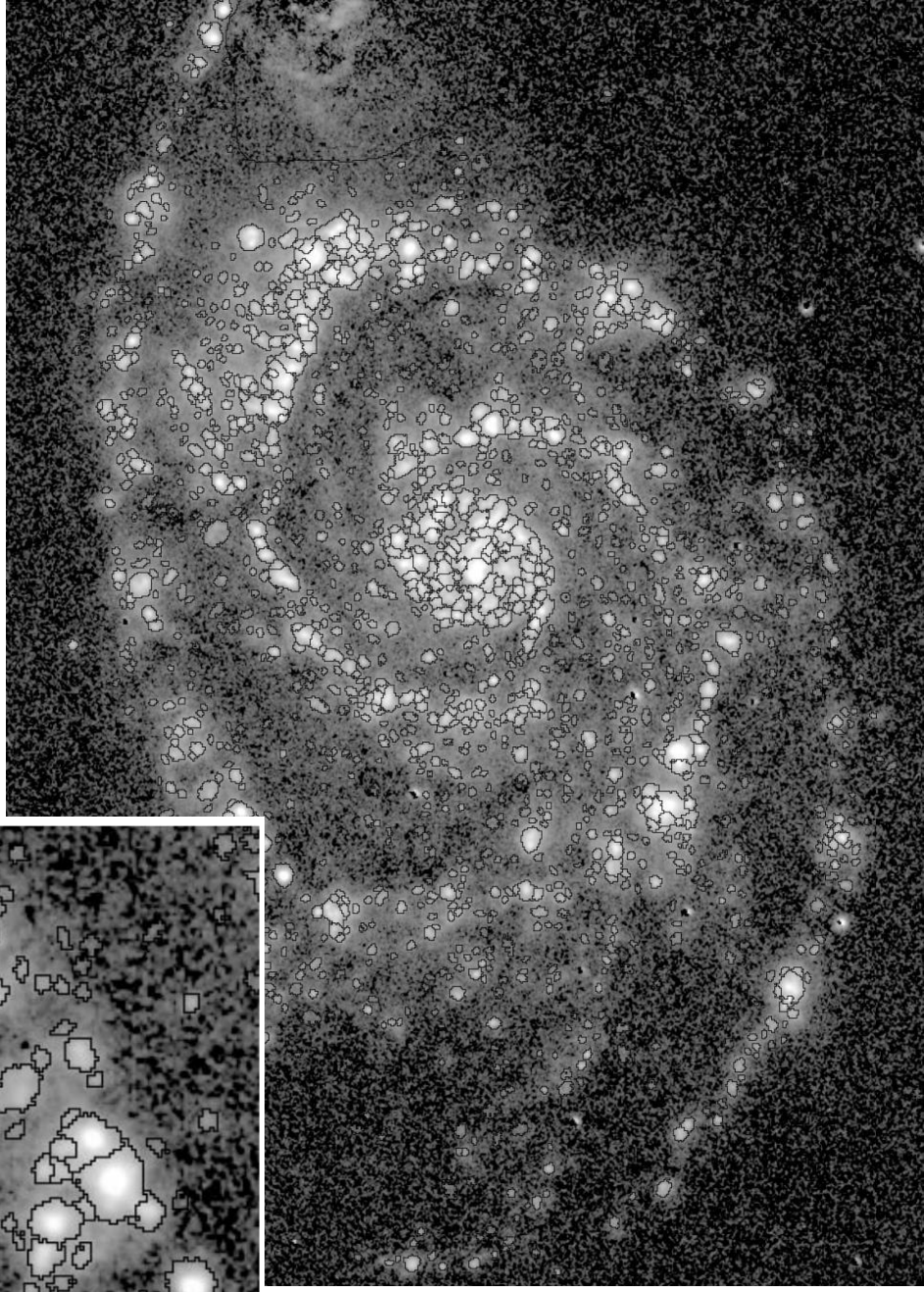
e.g. Salpeter Mass Function, source counts (number-magnitude counts from a deep image), primordial fluctuation spectrum, luminosity functions…

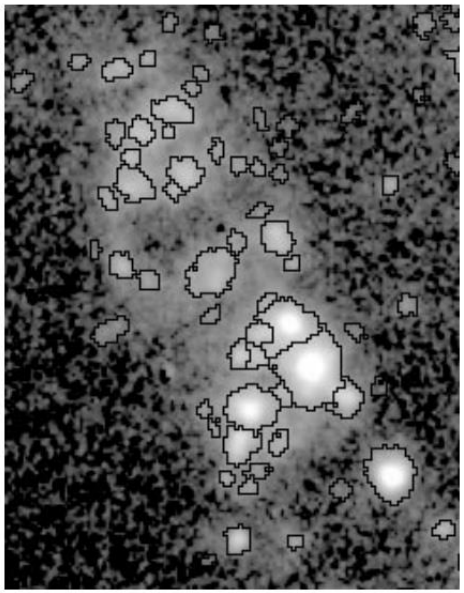
In its pure form it does not obey the formal definition of a pdf. But there are those physical limits that generally set upper and lower bounds.



HII region luminosity function (Liu et al. 2013)







Thilker (2001): HIIphot software

Many pitfalls of the power law! It has no saving grace via approximation to familiar well-bounded distributions.

a) Is this power law and *integral or differential* distribution? – a common way of getting the index wrong by 1. (**Rising or declining?**)

b) Is the *binning on a uniform or a log scale*? If its differential form is binned via a uniform $\Delta \log L$ scale instead of via ΔL, slope reduced by 1.

c) There is *no characteristic scale or spread* for such a distribution, although in practice the physical limits always provide high and low end-stops (Don't rely on them to make power laws tractable in terms of normal means and standard deviations).

Therefore, Wall & Jenkins comment:

Power-law distribution – the distribution from hell, because it does not conform, and our tacit (心照不宣)/ assumed reliance on Central Limit Theorem is lost. Many pitfalls to navigate regarding indices.