长尾分布，重尾分布(Heavy-tailed Distribution)

参考 长尾分布，重尾分布(Heavy-tailed Distribution) - 云+社区 - 腾讯云

Zipf分布：

Zipf分布是一种符合长尾的分布：

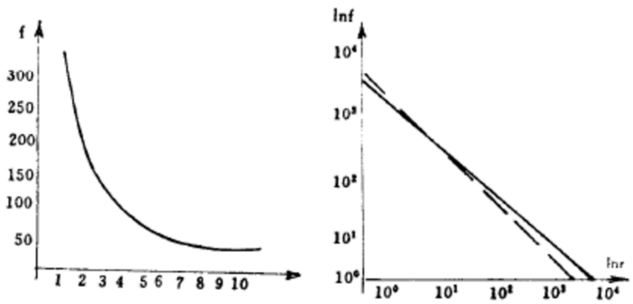
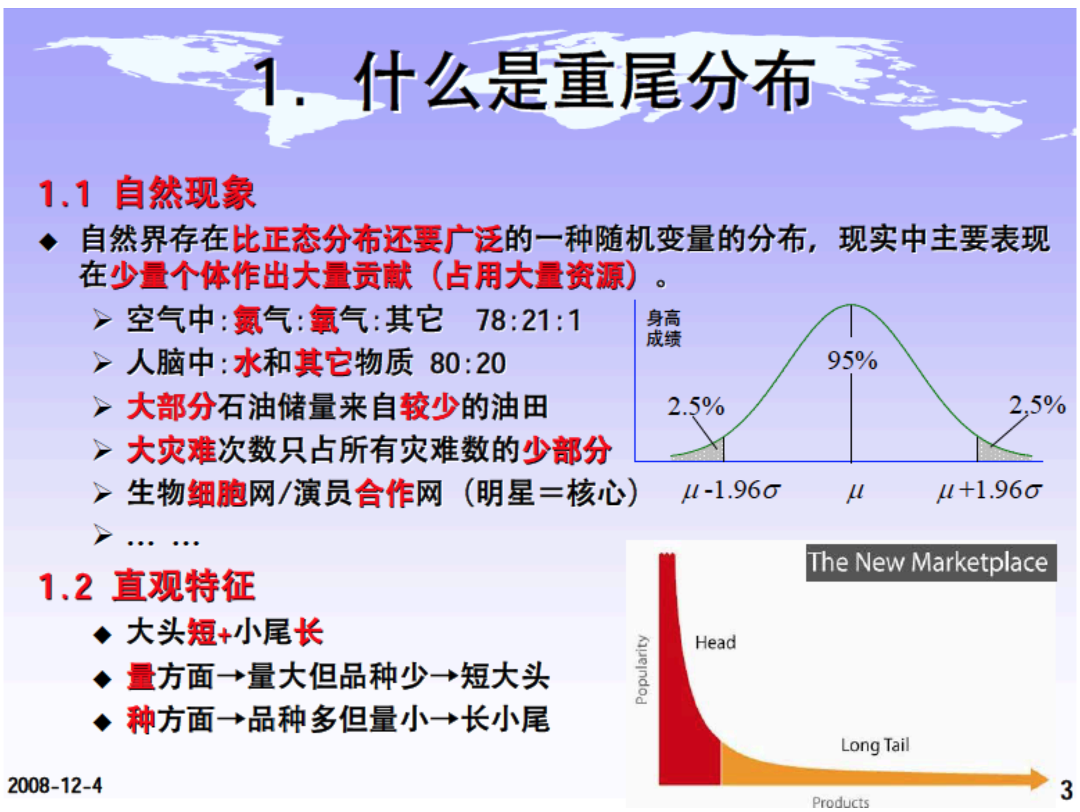


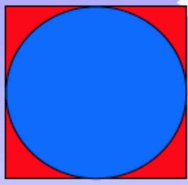
图 4—4 齐夫词频分布曲线

图 4—5 坐标轴为对数尺的齐夫分布曲线

就是指尾巴很长的分布。那么尾巴很长很厚的分布有什么特殊的呢？有两方面：**一方面，这种分布会使得你的采样不准，估值不准，因为尾部占了很大部分。另一方面，尾部的数据少，人们对它的了解就少，那么如果它是有害的，那么它的破坏力就非常大，因为人们对它的预防措施和经验比较少。也要所谓的二八法则。**



1.3 神秘的犹太大法则



$$\boxed{78}:\boxed{22}=\pi\left(\frac{1}{2}\right)^2:\left[1-\pi\left(\frac{1}{2}\right)^2\right]=\frac{\pi}{4}:\frac{4-\pi}{4}$$

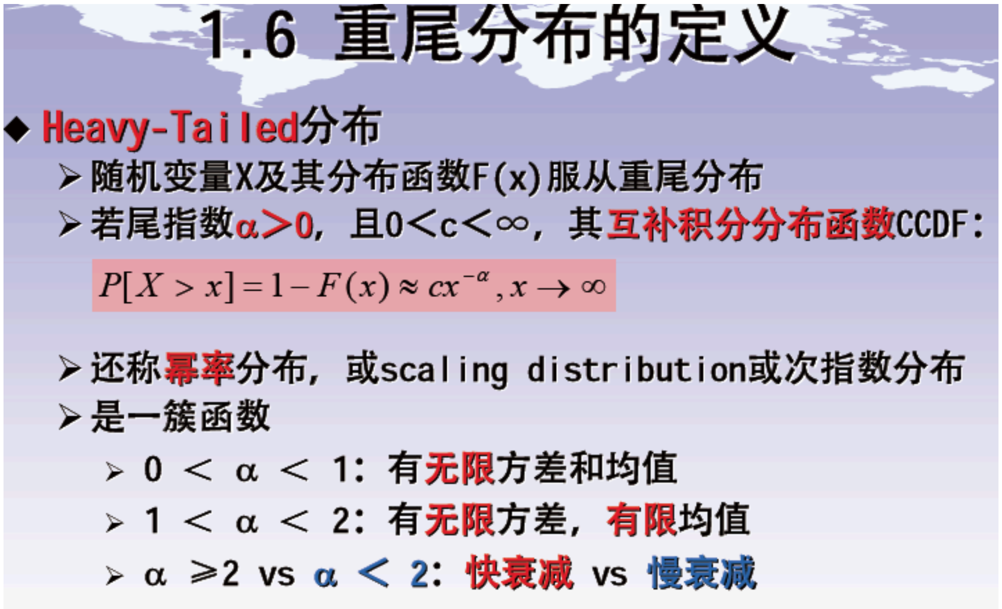
- Talmud：“**22：78是个永恒的法则，没有互让的余地。**”
  - 78%的**大众**占总财富的22%；22%的**富人**占总财富的78%；
  - 78%**贷**款人：22%**借**款人；
  - 78%**生意**来自22%**客户：**
    - “做生意永远只抓两件事情：女人和嘴巴”
  - 78%人**卖**时间；22%人**买**时间；
  - … …
- **奇妙比例、自然法则**
  - 人类**不可抗拒**，也是人类生存的**法则**。
  - 显然与**均匀分布**相距甚远

1.4 Pareto法则，或80/20法则

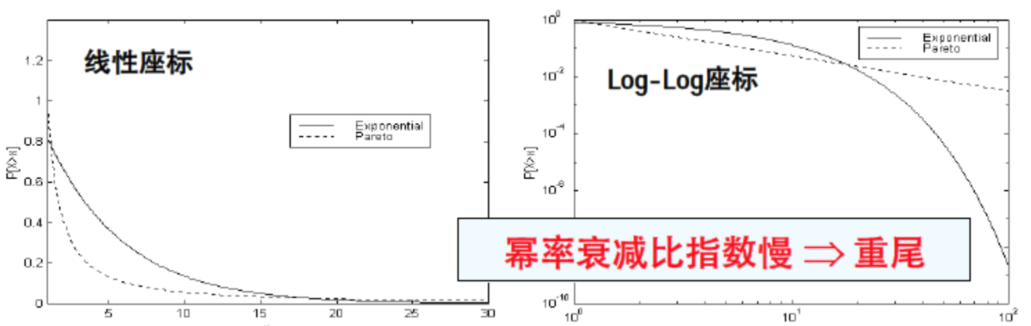
- 1906年经济学家Pareto观察众多现象后发现
  - 在任何一组东西之中，**最重要**的通常只占其中的一**小部分**
  - “**重要的少数**”和“**微不足道的多数**”
  - 80%的**结果**取决于20%的**原因**
- 现象
  - Italy **20%人口拥有80%财产**（1906）
  - 80%的劳动成果取决于20%的前期努力
  - 80%的**销量**来自与20%的**客户**。
  - 80%的**理赔额**支付给20%的**对象**
  - 个人在80%的**时间**里，穿着了自己20%的**服装??**… …

1.5 Zipf's 经验法则

- **只少数几个词汇经常被使用**;许多甚至大多数词汇**很少被用**
  - $P_r = c/(r+a)^b$; $0\le a<1$, $b>0$, $c>0$; $r=1,2\ldots$是某词汇出现频率的顺序
- 使用率超过80%的英文字母不到总数的50%
- **500**个常用汉字（500/2500=**20%**）覆盖率可达到**78%**，2500字达99.2%
  - 1994年《中华字海》收字85000。两级字库=3775+3008=6783:重尾分布


1.6 重尾分布的定义
◆ Heavy-Tailed分布
➢ 随机变量X及其分布函数F(x)服从重尾分布
➢ 若尾指数α>0，且0<c<∞，其互补积分分布函数CCDF：
$P[X > x] = 1 - F(x) \approx cx^{-\alpha}, x \to \infty$
➢ 还称幂率分布，或scaling distribution或次指数分布
➢ 是一簇函数
➢ 0 < α < 1：有无限方差和均值
➢ 1 < α < 2：有无限方差，有限均值
➢ α ⩾2 vs α < 2：快衰减 vs 慢衰减


Heavy-tailed 分布
◆ 亚指数分布的特例
◆ 渐进双曲线、幂率 (power-law) 形状
◆ CCDF: $1 - F(x) \sim x^{-\alpha} \quad 0 < \alpha \leq 2$
◆ 其 PDF 是幂率函数: $f(x) \sim x^{-\alpha-1}$
◆ 短尾与长尾的比较
– 点线：重尾Pareto分布，实线：轻尾指数分布


线性座标
Exponential
Pareto
Log-Log座标
Exponential
Pareto
幂率衰减比指数慢 ⇒ 重尾

1.7 常见的重尾分布

- **最简单的重尾分布**是**Pareto**分布（当常数c=k$^\alpha$）
  - 若X是一个随机变量， 则X的**Pareto** CCDF如下:
  - 1-F(x) = (k/x)$^\alpha$， 0＜α,x ＞k $P[X>x]=k^\alpha x^{-\alpha}$ when $x \ge k$
  - 概率密度函数 $f(x)=\alpha k^\alpha x^{-\alpha-1},\quad \alpha,k>0, x \ge k$
    - k决定随机变量可取的最小值，参数α决定随机变量均值和方差
    - 若将该CCDF取对数，其图形将表现为斜率为-α的直线
- 威伯 **Weibull**分布 $f(x)=\begin{cases} \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-(\frac{x}{\alpha})^\beta} & x>0, \text{常数}\beta>0, \alpha>0 \\ 0 & otherwise \end{cases}$
  - 若α＜ 1，则Weibull分布**属于重尾分布**;
- 对数正态**lognormal**分布: logX normally distr.,PDF

$$f(x)=\frac{1}{x\sqrt{2\pi\sigma}}\exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\},\ x>0$$

1.8 认识短尾与长尾

- 比较指数分布的**尾部形状** $1-F(x)=P[X>x]$ for large x
  - 指数分布：**更快**衰减，**短尾或轻尾** $1-F(x)\sim e^{-\lambda x}$
  - 亚指数分布：**较慢**衰减，**长尾或重尾**，尾部的观察**不可忽略**
- **重尾**分布的随机变量
  - 其**互补积分分布函数**（CCDF）曲线**衰减慢**于指数分布;
  - 意味者**相当大的概率质量**集中在**分布的尾部**。

$$\lim_{x\to+\infty}\frac{(1-F(x))}{e^{-\varepsilon x}}=+\infty\text{，对于某个}\varepsilon>0$$

  - **前n项部分和**与该前n项的**最大值是尾等价的**，反映了20%--80%现象

$$\lim_{x\to\infty}\frac{\overline{F^{(n)}}(x)}{\overline{F}(x)}=n\text{，}\quad\text{或等价地，}\quad\lim_{x\to\infty}\frac{P(X_1+X_2+\cdots+X_n>x)}{P(\max(X_1,X_2,\cdots,X_n)>x)}=1$$

Pareto与指数分布比较

- 缩放特性

$$P[X > x \mid X > w] = P[X > x] / P[X > w] \approx c_1 x^{-\alpha}$$

$$P[X > x \mid X > w] = \exp(-(x - w))$$

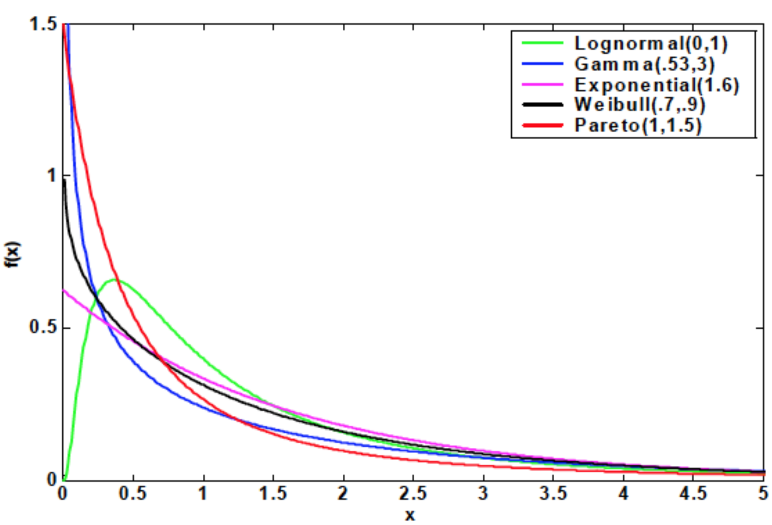
- 平均剩余寿命线性增长性

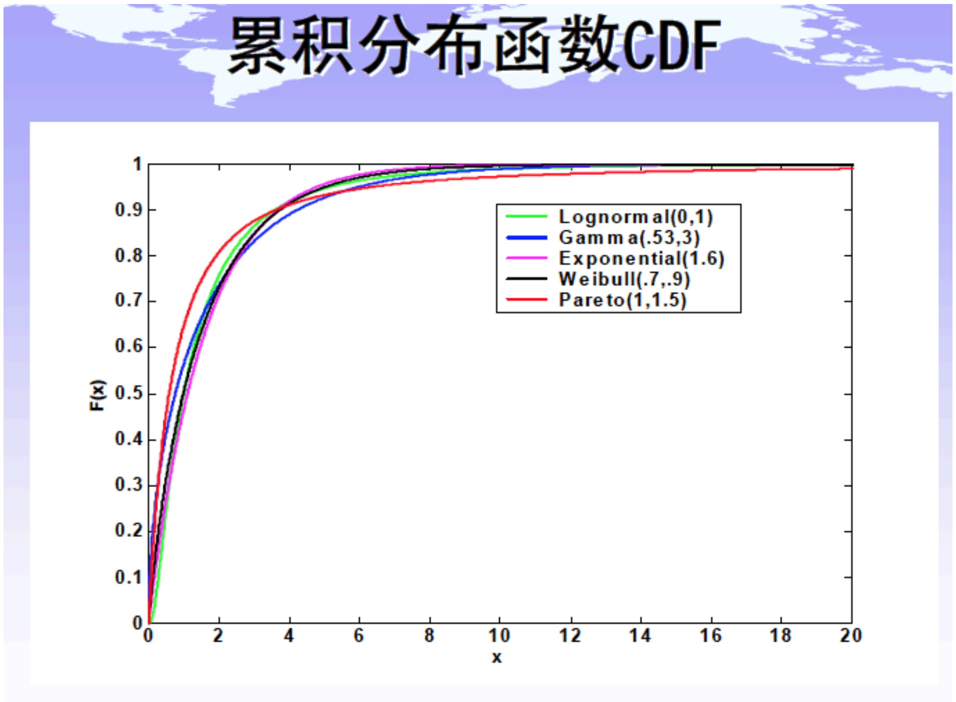
$$E[X - x \mid X > x] \approx cx$$ Long tail

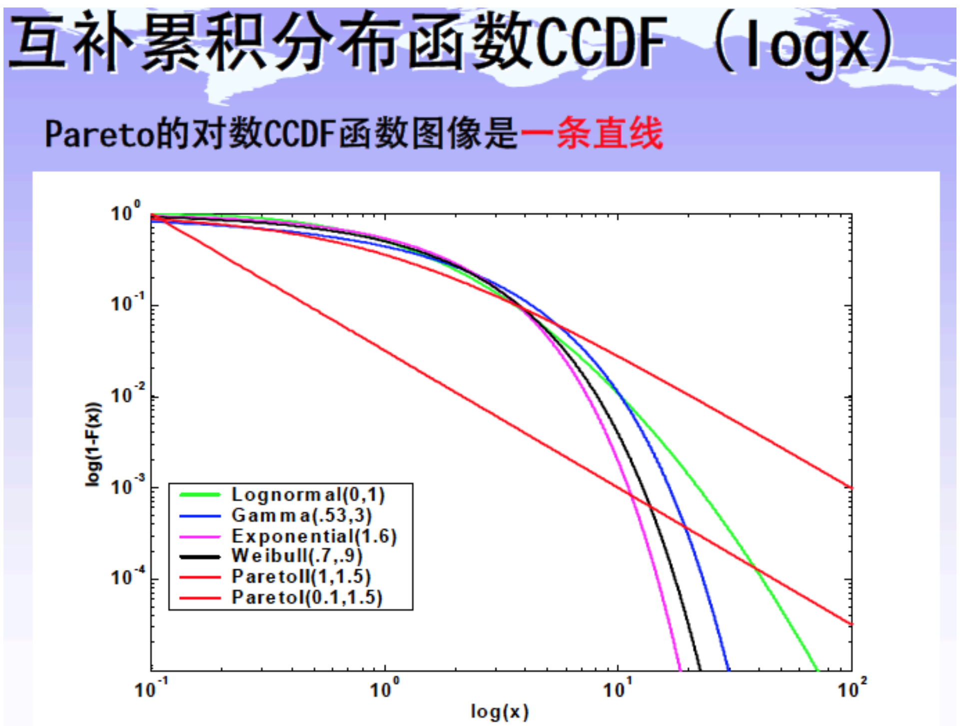
$$E[X - x \mid X > x] = const$$

- 不变性，对iid 1<α<2
  - 线性聚合后，前n项和的分布仍然是重尾分布；
  - 求最大仍是重尾分布
  - 加权混合后仍是重尾分布

重尾分布簇的概率密度函数PDF




累积分布函数CDF
Lognormal(0,1)
Gamma(.53,3)
Exponential(1.6)
Weibull(.7,.9)
Pareto(1,1.5)
F(x)
x


互补累积分布函数CCDF（logx）
Pareto的对数CCDF函数图像是一条直线
Lognormal(0,1)
Gamma(.53,3)
Exponential(1.6)
Weibull(.7,.9)
ParetoII(1,1.5)
ParetoI(0.1,1.5)
log(1-F(x))
log(x)