# 第六讲　爱上统计学

樊潇彦

复旦大学经济学院

# 本讲主要内容

## 1. 科学研究和实证分析

1.1 什么是科学？

1.2 实证分析从零假设开始

## 2. 概率统计基础

2.1 什么是概率？

2.2 随机变量与数字特征

2.3 大数定律和中心极限定理

## 3. 统计推断

3.1 什么是统计学？

3.2 样本与统计量

3.3 假设检验

## 4. 经济应用

4.1 汽车保险

4.2 资产收益率

# 从确定到随机：科学思想的演进

- 概率理论的所有认识论价值在于：大规模随机现象的集体行为产生严格的、非随机规律。

—— Gnedenko and Kolmogorov(1956)

- 机械式宇宙观认为，宇宙如同一个庞大的时钟机器，所有的物体都按照一定的规律运动，所有将来发生的事件都决定于过去的事件。.....今天，医学研究运用精巧的分布数学模型来确定治疗方法对长期存活的可能效果；社会学家和经济学家用数学分布来描述人类社会的行为；物理学家用数学分布来描述次原子粒子。科学里没有哪一个方面从这场革命中逃脱。有的科学家宣称，概率分布的使用只是一时的权宜之中，最终我们会找到一种途径回到19世纪科学的决定论。爱因斯坦有句名言，他不相信上帝在和宇宙玩骰子，就是这种观点的例子。其他人则相信，大自然基本上是随机的，真实性只存在于分布函数之中。

—— D. 萨尔斯伯格，《女士品茶》

# 什么是科学？

- 设想上帝正在玩某种伟大的游戏，比如下棋。你不懂这种游戏的规则，但允许你在场上看，至少可以不时地在一个小角落里观棋。通过这些观察，你试图搞明白这游戏的规则是什么，走棋的规则是什么。
- 目前而言，我们只有通过数学形式的推理才能了解这世界的最终特性。……不懂数学对于你理解这世界来说就是个严重缺陷。
- 因为科学的成功，所以就有了伪科学。社会科学就是一个例子，它就是那种不是科学的科学。他们不是科学地做事情，而是徒有科学研究的形式——或者，收集数据，做这做那，但他们得不出任何规律，他们没发现什么。……（专家的论断）也许对，也许不对，不过至今还没有任何途径来证实它。

—— R.P. 费曼，《发现的乐趣》

好奇 + 观察 + 推理 + 检验 = 科学

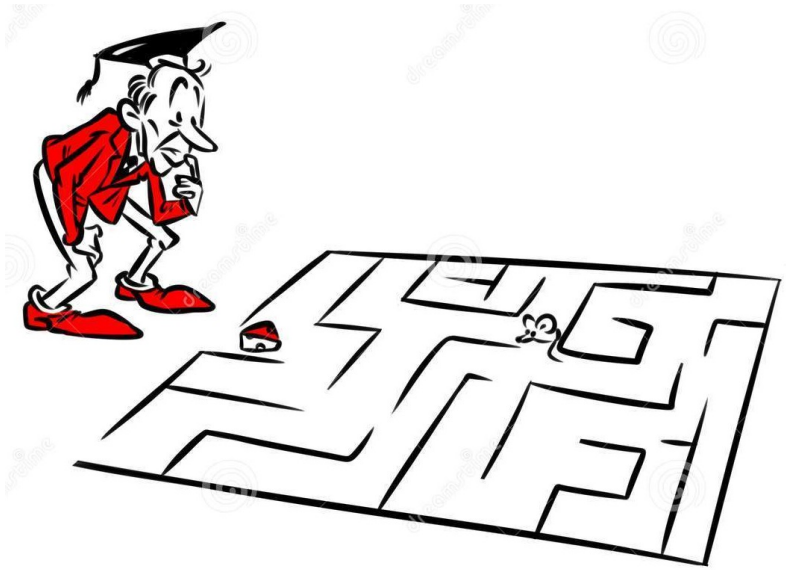
# 科学研究的步骤

1. 一个简要、清晰的问题；
2. 通过逻辑思考和数理分析，得到一个可供证伪的理论假说（theoretical hypothesis）；
3. 根据理论假说设计实证研究方案，借助于逻辑思考和统计方法，构造与理论假说相对应的实证假设（empirical hypothesis）；
4. 通过数据采集、整理、统计和计量分析，检验实证（和理论）假说，回答所关心的问题。

# 科学研究的注意事项

- 如果只关心“是什么”的问题，可以跳过步骤2，但如果要回答“为什么”的问题，则必须根据理论假说构造实证假设；
- 实证研究的关键在于设计实验方案；
- 数据采集应满足问题相关、代表总体和样本充足3方面的要求；
- 在样本量很大的情况下，需要用到较新的统计和计量分析技术，如统计学习、非参估计等。

# 例：费曼先生眼中的“甲等一号实验”



老鼠的记忆：

- 气味？
- 视觉？
- 声音？

从一个科学的立场来看，这是个甲等一号实验。

- 因为它揭示了老鼠实际使用的线索，而不是你认为老鼠使用的线索。
- 这个实验准确地告诉你，为了在一个跑老鼠实验中能够留心并掌控每一件事，你必须运用哪些条件。

# 实证分析从零假设开始

- 零假设（Null hypothesis）就是无差异或无关陈述，它既是研究的起点也是判定的基准，例如：
  - A和B没有差别；
  - A对B没有影响；
  - 即使在条件A下，B对C也没有影响。
- 零假设就像“无罪推定”，只有提供充分的证据才能证明一个人有罪，否则就只能接受它。科学研究要让人们相信一个结论，也要提供充分的证据拒绝零假设。
- 什么是好的实证假设？
  - 以陈述句的形式表述假定变量间的关系
  - 反映假设建立的理论和机制
  - 简短、切中要点、可检验

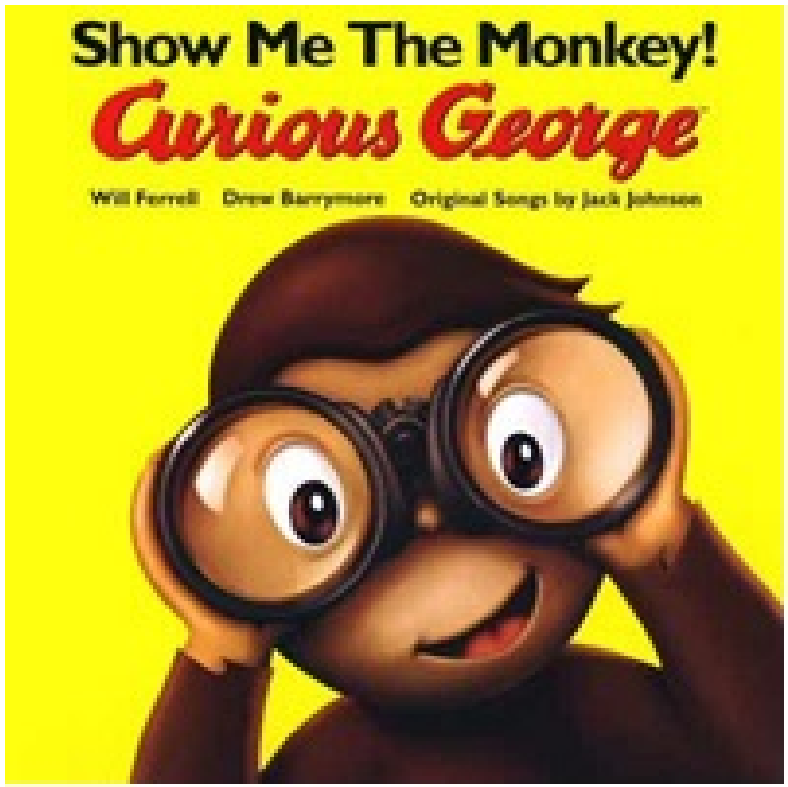
# 例：什么是好的实证假设？

在20 世纪之交，统计学家 K. 皮尔逊和其他学者创办了《生物统计》(Biometrika)，他们做的研究包括：

- 收集生物样本数据，检验达尔文环境影响生物进化的假说；
- 分析从世界各地收集到的犹太人与非犹太人的人体测量数据，最后得出的结论是：纳粹的种族理论纯粹是胡说八道，根本就没有犹太种族（Jewish race）或亚利安种族（Aryan race）那回事。
- 对澳洲原住民的人类学测量与对欧洲人的测量结果有着相同的分布，据此推翻了某些澳洲人关于原住民不是人类的断言。

—— 萨尔斯伯格，《女士喝茶》

# 《爱上统计学》从一个问题开始...

# 古典概型与几何概型

- 古典概型：（1）样本空间 $\Omega = \{\omega_1, \omega_2, ...\omega_N\}$ 只有有限个样本点；（2）每个样本点出现的可能性相同。以掷骰子为例，掷出某个或某几个点数的随机事件 $A$ 出现的概率：

$$P(A) = \frac{A\text{中的样本点数}}{\Omega\text{中的样本点数}}$$

- 几何概型：（1）样本空间 $\Omega$ 中有无限个样本点；（2）每个样本点出现的可能性相同。例如 $\Omega = [0, 1]$，$A = [0, 0.2]$，则 $A$ 出现的概率：

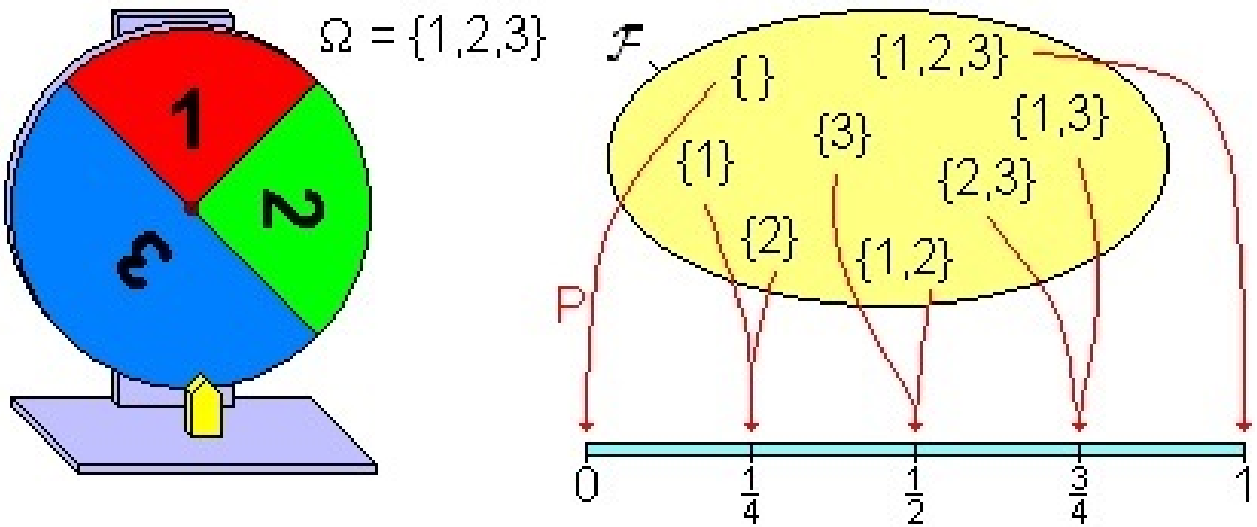
$$P(A) = \frac{A\text{中线段的长度}}{\Omega\text{中线段的长度}}$$

# 概率的公理化定义

根据柯尔莫哥洛夫（Kolmogorov，1933），若函数 $P:\mathscr{F}\to[0,1]$ 满足：

- 非负性 $P(A)\geq 0$；
- 规一性 $P(\Omega)=1$；
- 可列可加性，即如果 $A_i\in\mathscr{F}$ 且 $A_iA_j=\emptyset,(i\neq j)$

则称 $P$ 为可测空间 $(\Omega,\mathscr{F})$ 上的一个概率测度（probability measure），简称概率。




Source: “Probability measure” on Wiki.

# 条件概率与三个重要公式

$(\Omega, \mathscr{F}, P)$ 为概率空间，$A, B \in \mathscr{F}$，且 $P(B) > 0$，定义 $P(A|B)$ 为事件 $B$ 发生下事件 $A$ 发生的条件概率（conditional probability）：

$$P(A|B) = \frac{P(AB)}{P(B)}$$

- 乘法公式：$P(AB) = P(B)\,P(A|B)$
- 全概率公式：若 $\{B_i,\ 1 \le i \le N\}$ 为样本空间 $\Omega$ 的一个分解，则：

$$P(A) = \sum_{i=1}^{N} P(B_i)\,P(A|B_i)$$

- 贝叶斯定理（Bayes' theorem）：

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

# 例：选股必涨的巴尔的摩股票经纪人

有一天，一位巴尔的摩的股票经纪人主动给你发来一份行业资讯，透露了某只股票将要大涨的内幕消息。一周之后，这位巴尔的摩股票经纪人的预言应验了，这支股票真的涨了。第二周你又收到一期行业资讯。这一次，这位经纪人认为某只股票会跌，结果这支股票真的跌了。10周过去了，这份神秘的行业资讯每期都有新预测，而且他们全都应验了。

第11周，你会让这位巴尔的摩股票经纪人帮你做投资吗？

—— J. 艾伦伯格，《魔鬼数学：大数据时代数学思维的力量》

# 例：你的邻居是恐怖分子吗？

假如有一天，你发现邻居的名字出现在脸谱发布的“恐怖分子嫌疑人”的名单上，而且你觉得脸谱的2亿客户有如下分布：

| | 出现在名单中 | 没有出现在名单中 |
|---|---|---|
| 是恐怖分子 | 10 | 9,990 |
| 不是恐怖分子 | 99,990 | 199,890,010 |

你觉得你的邻居是恐怖分子吗？

—— J. 艾伦伯格，《魔鬼数学：大数据时代数学思维的力量》

# 例：911是恐怖袭击吗？

我们用 $C_1$, $C_2$ 表示第一次和第二次飞机撞击，$T$, $NT$ 表示恐怖袭击和意外事故，发生概率如下：

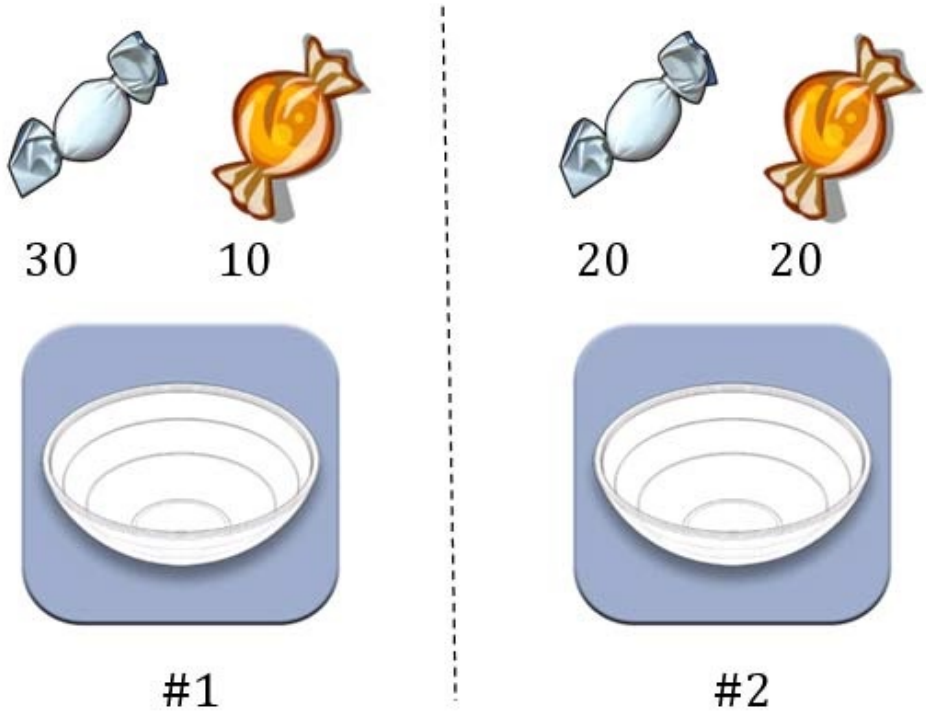
| | 概率 | 含义 |
|---|---|---|
| $P(T)$ | 0.005% | 美国遭受恐怖袭击的先验概率 |
| $P(C_1\|T) = P(C_2\|T)$ | 1 | 恐怖分子驾机撞上了世贸中心大楼 |
| $P(C_1\|NT) = P(C_2\|NT)$ | 0.008% | 飞机意外撞上了世贸中心大楼 |

请计算在第一次和第二次飞机撞击世贸中心大楼之后，美国遭受恐怖袭击的后验概率分别为多少？

—— N. 西尔弗，《信号与噪声》

# 例：来自哪只碗？

两个一模一样的碗，一号碗有30颗水果糖和10颗巧克力糖，二号碗有水果糖和巧克力糖各20颗。现在随机选择一个碗，从中摸出一颗糖，发现是水果糖。请问这颗水果糖来自一号碗的概率有多大？




资料来源：阮一峰，“贝叶斯推断及其互联网应用（一）：定理简介”

# 随机变量

$(\Omega, \mathscr{F}, P)$ 为一个概率空间，若定义在样本空间上的实值函数 $X : \Omega \to \mathbb{R}$ 满足 $\forall x \in \mathbb{R}$ 有：

$$\{\omega : X(\omega) \leq x\} \in \mathscr{F}$$

则称 $X$ 为 $(\Omega, \mathscr{F}, P)$ 上的随机变量。

- 离散型随机变量：二项分布（伯努利分布）、泊松分布等
- 连续型随机变量：均匀分布、正态分布（高斯分布）、指数分布，以及$t$、$F$、$\chi^2$ 分布等

# 随机变量的累积分布和概率密度函数

- $X$ 是概率空间 $(\Omega, \mathscr{F}, P)$ 上的随机变量，若对于 $\forall x \in \mathbb{R}$ 存在：

$$F(x) = P(X \leq x)$$

称 $F : \mathbb{R} \to [0,1]$ 为累积分布函数（cumulative distribution function，cdf）。

- 对连续随机变量 $X$，如果除有限个点外存在：

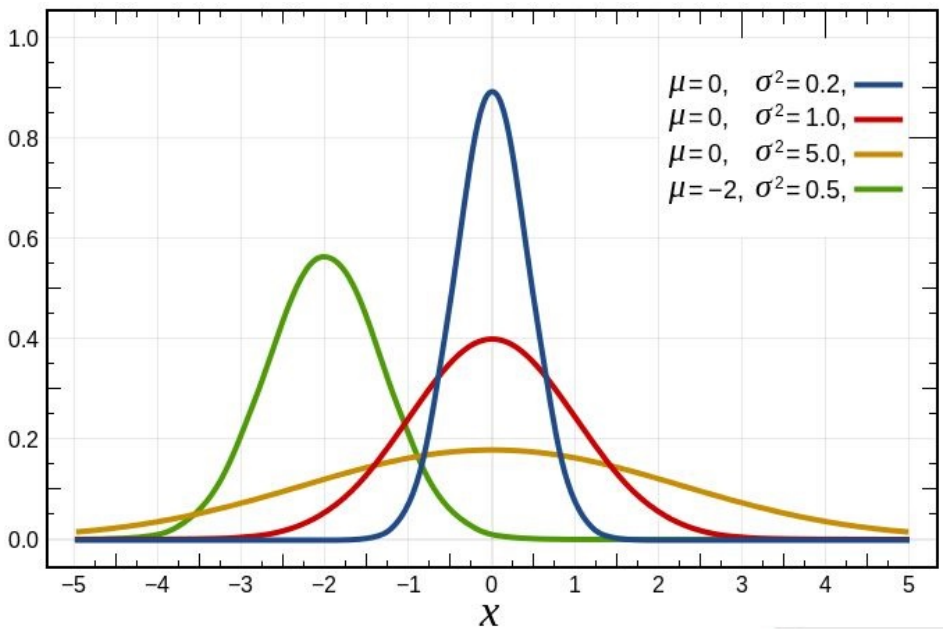
$$f(x) = \frac{dF}{dx}, \qquad \int_{-\infty}^{x} f(u)\,du = P(X \leq x) = F(x)$$

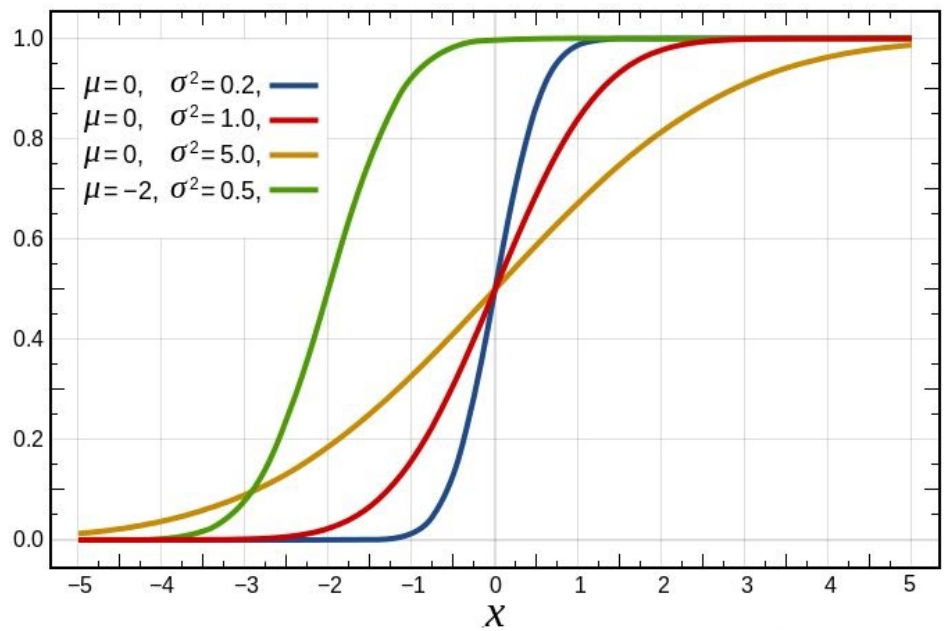
称 $f : \mathbb{R} \to [0, \infty)$ 为概率密度函数（probability density function，pdf）。

- 对于离散随机变量 $X \in \{x_i\}_{i=1}^{N}$，称 $p(x_i) = P(X = x_i)$ 为频率函数（frequency function）或分布律。

# 正态分布 $x \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$




图片来源：Wiki百科的正态分布的概率密度（左）和累积分布（右）函数。

# 随机变量的数字特征：数学期望和矩

- 对于 $\mathbb{R}$ 上的函数 $g(x)$ 和累积分布函数 $F(x)$，定义 $g(x)$ 的数学期望（mathematical expectation）为：

$$E[g(x)] = \int_{-\infty}^{\infty} g(x) dF(x)$$

- 对任意 $k, l \in \{1, 2...\}$，定义随机变量的矩（moment）：
  - 若 $E\left(x^k\right)$ 存在，称之为 $X$ 的 $k$ 阶原点矩，均值 $E(x)$ 为一阶原点矩；
  - 若 $E\left[(x - E(x))^k\right]$ 存在，称之为 $X$ 的 $k$ 阶中心矩，方差 $Var(x)$ 为二阶中心矩；
  - 若 $E\left[(x - E(x))^k (y - E(y))^l\right]$ 存在，称之为 $X$ 和 $Y$ 的 $k + l$ 阶混合中心矩，协方差 $Cov(x, y)$ 是二阶混合中心矩。

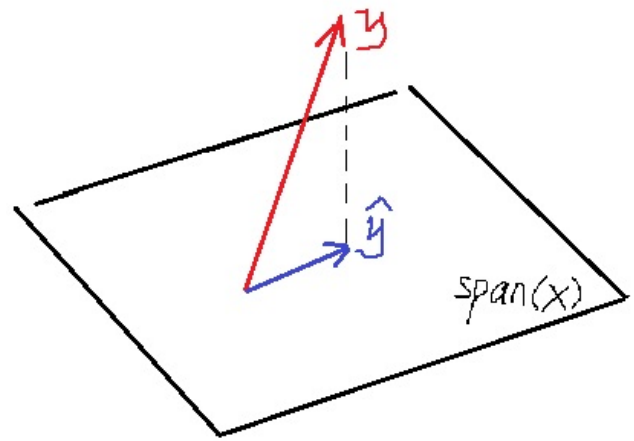
# 随机变量的数字特征：条件数学期望与线性投影

- 随机变量 $X$ 和 $Y$，如果 $\int_{-\infty}^{\infty} ydF(y|x)$ 存在，则称之为 $Y$ 关于 $X=x$ 的条件数学期望（conditional expectation）：

$$E(y|x)=\int_{-\infty}^{\infty} ydF(y|x)$$

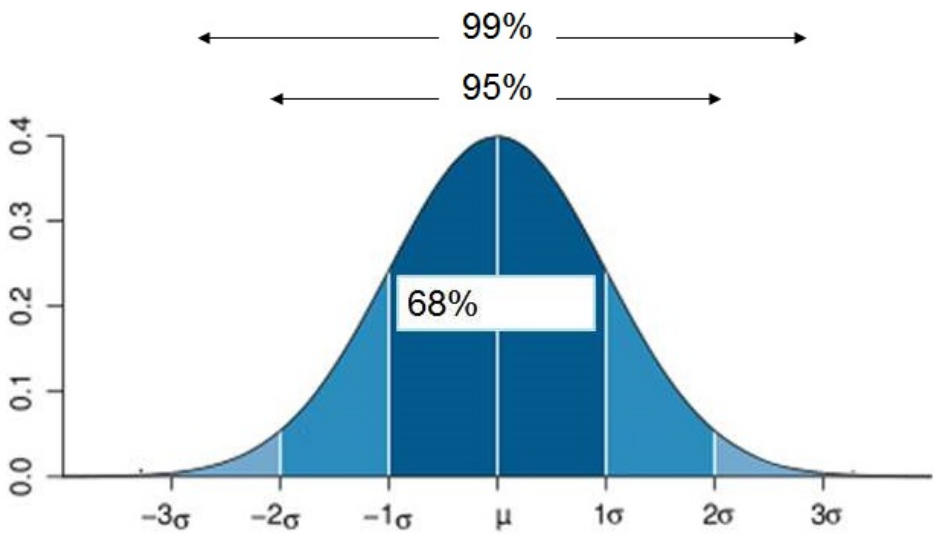
- 给定矩阵 $X$，我们想找到平面 $span(X)$ 中和 $y$ 的距离最短的向量 $\hat{y}$，使 $\varepsilon=y-\hat{y}$ 的长度 $\|\varepsilon\|$ 最短。我们称 $\hat{y}=X\beta$ 为 $y$ 在 $span(X)$ 上的线性投影（projection）。也就是说：

$$y=X\beta+\varepsilon \Rightarrow \hat{y}=E(y|X)=X\beta$$



# 例：正态分布 $N(\mu, \sigma^2)$ 的数字特征

- 均值（mean）：$\mu$，等于中位数和众数
- 方差（variance）：$\sigma^2$
- 偏度（skewness）：$S = 0$
- 峰度（kurtosis）：$K = 3$


图片来源：Wiki百科。

# 大数定律

研究在什么条件下，随机变量的均值收敛到一个常数。

- 定义随机序列 $\{\xi_n\}$ 依概率收敛于 $a$：存在常数 $a$，$\forall \varepsilon > 0$，有 $\lim\limits_{n\to\infty} P\left(|\xi_n - a| < \varepsilon\right) = 1$。
- 切贝谢夫定理：当随机变量序列 $\{x_i\}$ 独立分布，且方差有限 $Var\left(x_i\right) = \sigma_i^2 < \infty$ 时，前 $n$ 项算术平均 $\bar{x}_n$ 依概率收敛到其数学期望的均值 $\mu$：

$$\bar{x}_n \xrightarrow{p} \mu \qquad \text{其中：} \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \mu = \frac{1}{n}\sum_{i=1}^{n} E\left(x_i\right)$$

# 中心极限定理

研究在什么条件下，随机变量之和的分布收敛到正态分布。

- 假定随机变量序列 $\{x_i\}$ 独立分布，且方差有限 $Var(x_i) = \sigma_i^2 < \infty$ 时，定义随机变量：

$$\tilde{x}_n = \frac{1}{S_n}\sum_{i=1}^{n}[x_i - E(x_i)], \qquad S_n = \sqrt{\sum_{i=1}^{n}\sigma_i^2}$$

- 李雅普诺夫定理：当 $n \to \infty$ 时，上述随机变量 $\tilde{x}_n$ 的分布收敛到标准正态分布：

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P(\tilde{x}_n \le x) = \Phi(x) = \frac{1}{2\pi}\int_{-\infty}^{x} e^{-t^2/2}dt$$

记为 $F_n(x) \xrightarrow{w} \Phi(x)$，此时称 $\tilde{x}_n$ 依分布收敛（弱收敛）于标准正态随机变量 $\tilde{x}_n \xrightarrow{d} x$。

# 统计学

统计学（statistics）是描述一系列可用于描述、解释和分析资料或数据的工具和技术。

- 描述（descriptive）统计：描述样本（sample）数据的特征；
- 推断（inferential）统计：基于样本数据推断总体（population）数据的特征。

# 样本与统计量

- $\{x_i\}_{i=1}^N$ 是随机变量 $X$ 的 $N$ 个独立抽取的样本，与总体矩相对应的统计量（statistic）定义如下：

| | | | |
|---|---|---|---|
| 总体均值 | $\mu = E(x)$ | 样本均值 | $\bar{x} = \frac{1}{N}\sum_{i=1}^N x_i$ |
| 总体方差 | $\sigma^2 = Var(x)$ | 样本方差 | $s^2 = \frac{1}{N-1}\sum_{i=1}^N (x_i - \bar{x})^2$ |
| 总体 $k$ 阶原点矩 | $E\left(x^k\right)$ | 样本 $k$ 阶原点矩 | $A_k = \frac{1}{N}\sum_{i=1}^N x_i^k$ |
| 总体 $k$ 阶中心矩 | $E\left((x-\mu)^k\right)$ | 样本 $k$ 阶中心矩 | $M_k = \frac{1}{N}\sum_{i=1}^N (x_i - \bar{x})^k$ |

- 注意：
  - 统计量是样本的函数，样本不同统计量也不同，因此统计量本身也是一个随机变量；
  - 统计量的分布称为抽样分布（sample distribution），它与随机变量 $X$ 所服从的总体分布（population distribution，$F(x)$）是两个不同的概念。比如根据中心极限定理，不论 $X$ 服从哪种分布，在一定条件下，样本均值 $\bar{x}$ 都将渐近服从正态分布 $N\left(\mu, \sigma^2/N\right)$。

# 常用统计量

- 集中性指标：

  均值 $\bar{x} = x_{mean} = E(x)$、中位数 $x_{med}$ 和众数 $x_{mode}$

- 变异性指标：

  极差$\sigma_{ext} = \max(x) - \min(x)$、方差 $\sigma^2 = Var(x) = E(x - \bar{x})^2$ 和标准差 $\sigma = sd(x)$

- 相关性指标：

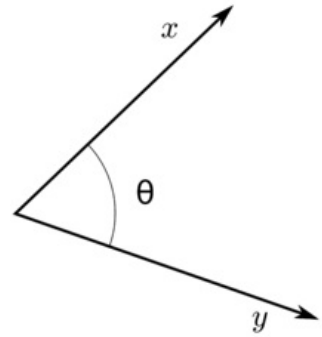
  - 协方差（covariance）：$Cov(x, y) = E(x - \bar{x})E(y - \bar{y})$
  - 相关系数（correlatin coefficient）：$Cor(x, y) = \frac{Cov(x,y)}{sd(x)sd(y)}$

# 相关系数的几何解释

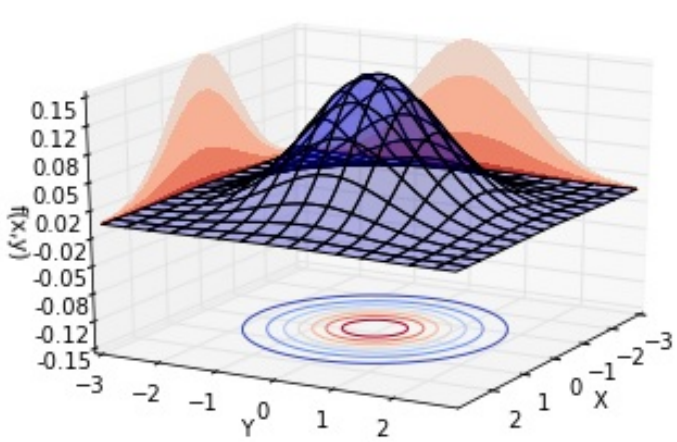
- 向量 $x, y$ 的相关系数（correlation coefficient）与两者之间夹角的关系为：

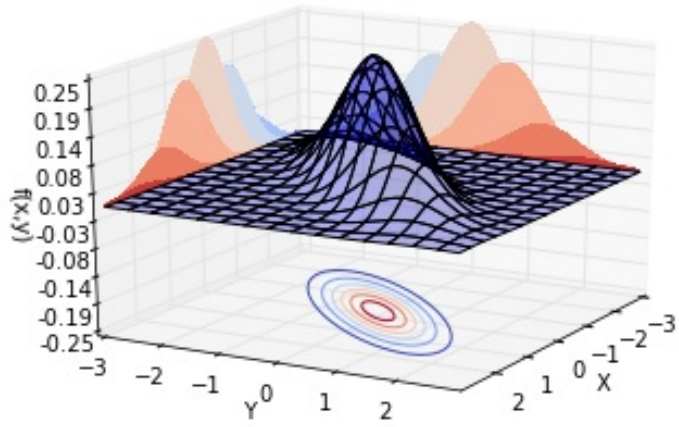
$$\rho = Cor(x, y) = \frac{x \cdot y}{\|x\| \|y\|} = cos(\theta)$$

两者独立时 $\rho = 0 \Leftrightarrow x \perp y$。




- 假定 $x$ 和 $y$ 服从正态分布，当 $\rho = 0$ 和 $\rho > 0$ 时，有：




图片来源：http://www.cnblogs.com/vamei/p/3416138.html

# 例：分析体验数据

请根据下面的体验数据，计算身高和体重的均值、方差和相关系数。

| | 160cm | 170cm | 180cm |
|---|---|---|---|
| 60kg | 0.2 | 0.05 | 0.05 |
| 70kg | 0.05 | 0.3 | 0.05 |
| 80kg | 0.05 | 0.05 | 0.2 |

# 假设检验

- 定义：根据样本判断关于总体分布 $X \sim F(x)$ 的假设是否正确。
- 分类：
  - 参数检验：给定总体分布的类型，检验对分布参数的假定是否正确

$$H_0 : \theta \in \Theta$$

  - 分布检验：检验对总体分布类型的假定是否正确

$$H_0 : X \sim F(x)$$

# 假设检验的两类错误

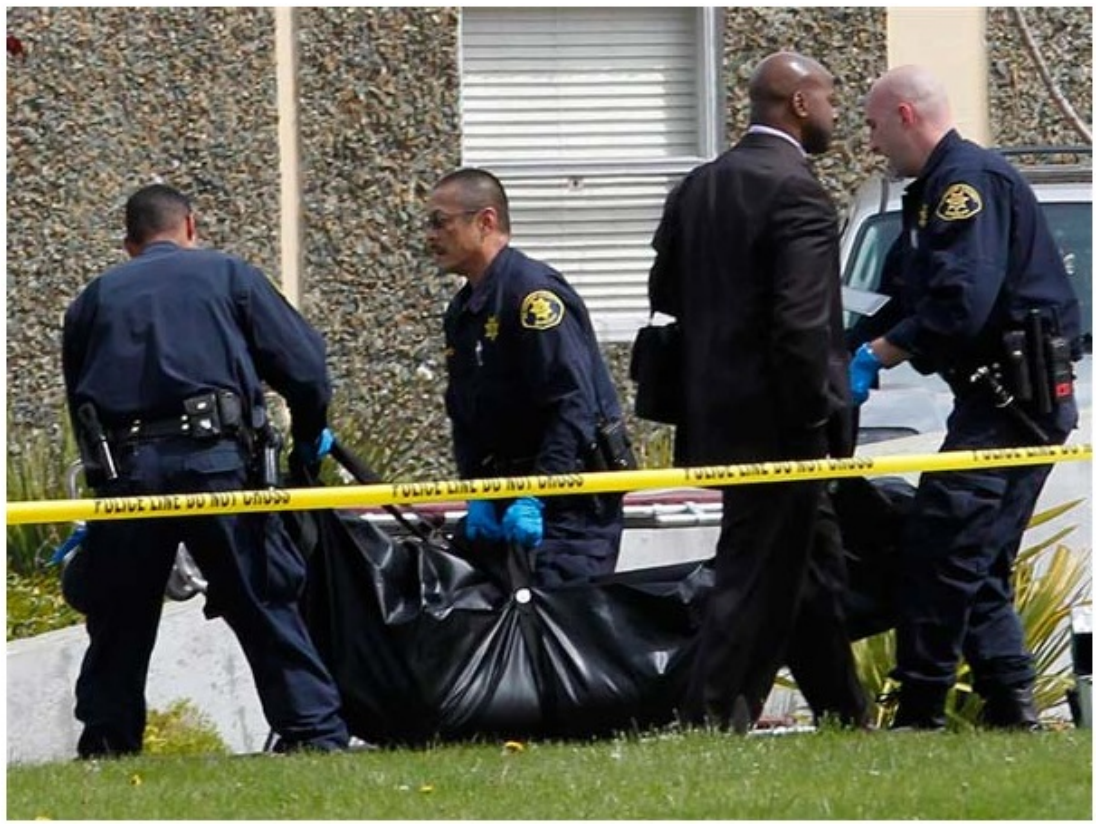
- 第一类错误（弃真）：$\alpha = P(\text{否定}H_0 | H_0\text{正确})$、犯第一类错误的概率 $p = 1 - \alpha$
- 第二类错误（取伪）：$\beta = P(\text{接受}H_0 | H_0\text{错误})$、检验的势 $\pi = 1 - \beta$

| | | 可能的选择 | |
|---|---|---|---|
| | | 接受零假设 | 拒绝零假设 |
| 零假设的真实性质 | 零假设是真实的 | 1<br>☺对啦，零假设是真实的情况下你接受了零假设，而且群体之间没有差别。 | 2<br>哎—你犯了第一类错误，在群体之间没有差异的情况下拒绝了零假设。第一类错误也可以用希腊字母阿拉法，或 α 表示。 |
| | 零假设是虚假的 | 3<br>哦—你犯了第二类错误，接受了虚假的零假设。第二类错误也可以用希腊字母贝塔，或 β 表示。 | 4<br>☺很好，在群体之间存在差异的情况下你拒绝了零假设。也可以叫做检定力，或 1 − β。 |

资料来源：N.J. 萨尔金德著，《爱上统计学》，史玲玲译，重庆大学出版社，2008

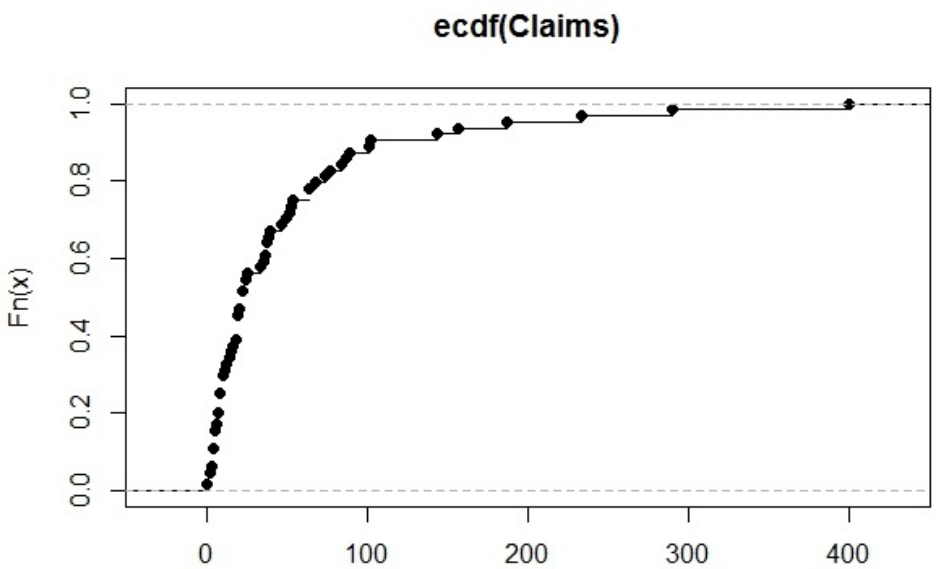
# 例：泊松分布与美国枪击案

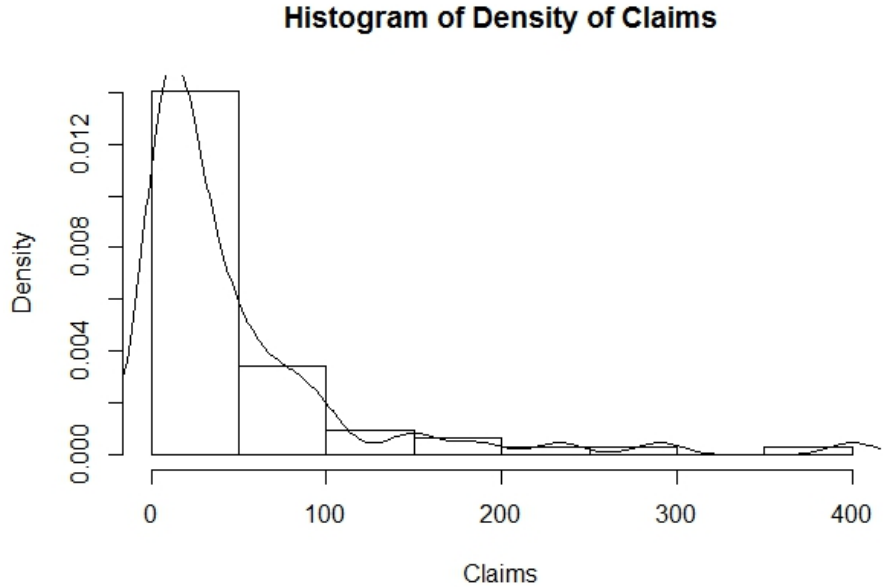
1982年至2012年，美国共发生62起（大规模）枪击案。其中，2012年发生了7起，是次数最多的一年。这是巧合，还是表明美国治安恶化了？



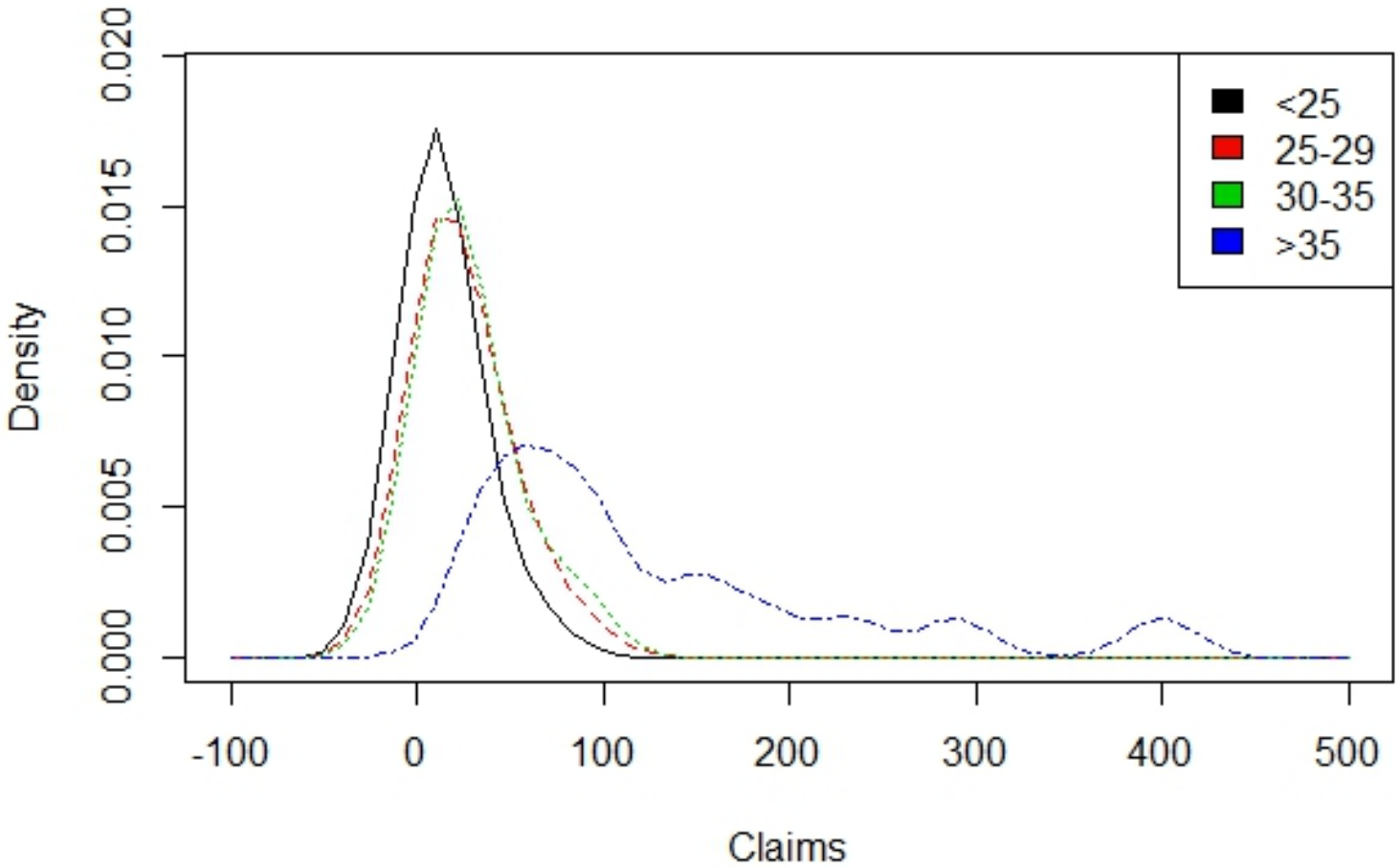

资料来源：阮一峰同名博客文章。

# 理赔额（Claims）的累积分布和概率密度函数

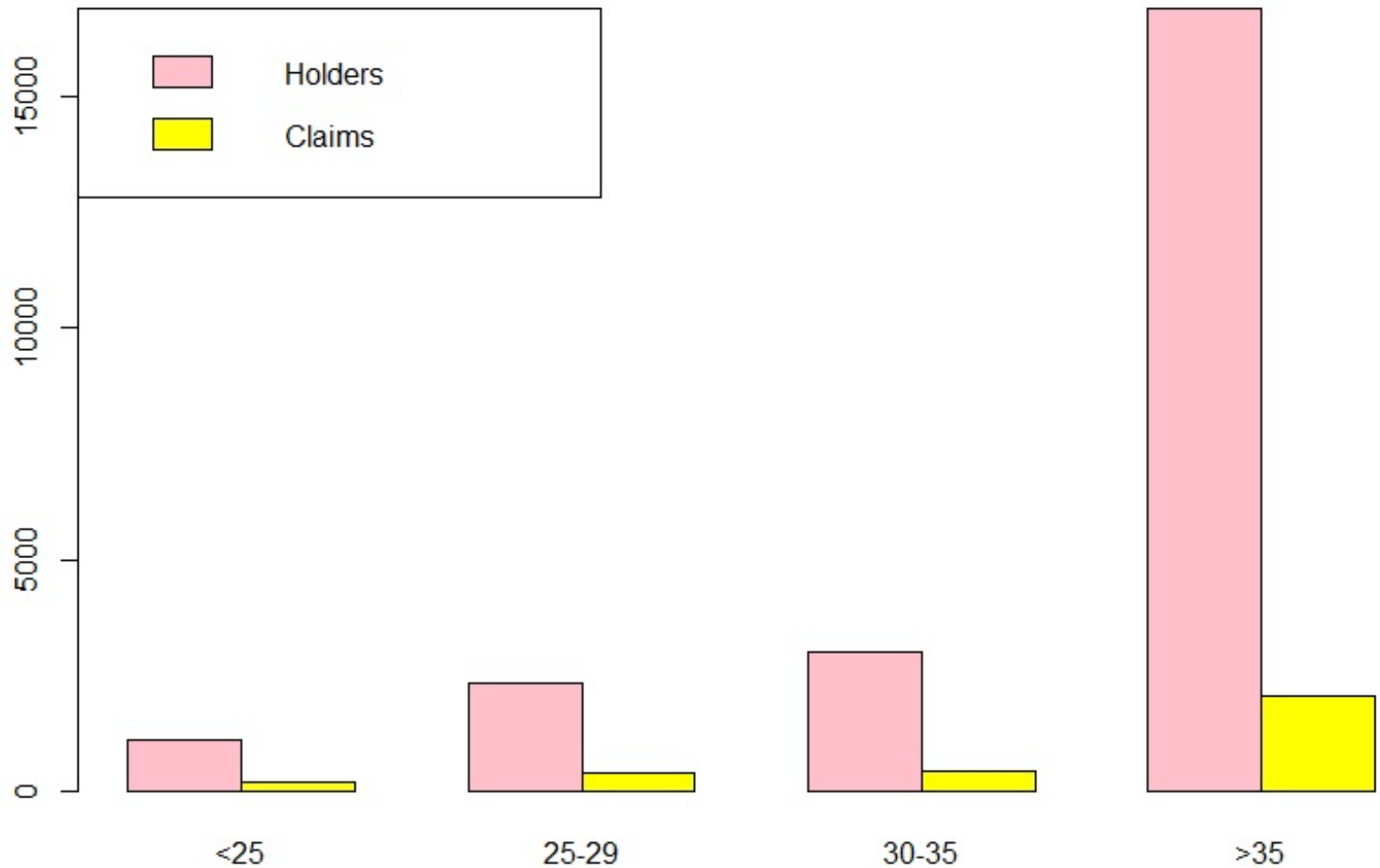


# 不同年龄组理赔额的概率密度函数



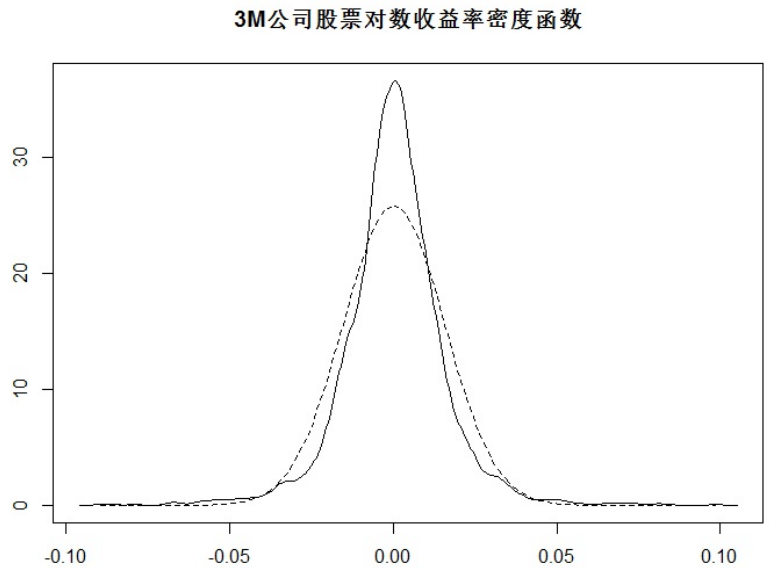
# 不同年龄组投保额（Holders）和理赔额的柱状图



# 资产收益率的统计特征

$$r_t = \mu + \rho r_{t-1} + \varepsilon_t$$
$$E(r_t) = \frac{\mu}{1-\rho}, \quad Var(r_t) = \frac{\sigma_\varepsilon^2}{1-\rho^2}$$

根据 Fan and Yao(2015) 资产收益率的典型特征包括：




- 平稳性(stationarity)：有稳定的均值和有限的方差
- 厚尾性(heavy tails)：峰度 $\hat{K} > 3$，与正态分布相比呈现出尖峰厚尾的特征，说明出现爆涨和爆跌的概率大于正态分布。
- 非对称性(asymmetry)：分布存在负偏（negatively skewed），说明市场下跌比上涨的程度大。
- 加总高斯性(aggregational Gaussianity)：当时间跨度上升时，相应的收益率会趋向正态分布，如年收益率与月收益率和日收益率相比，更接近于正态分布。