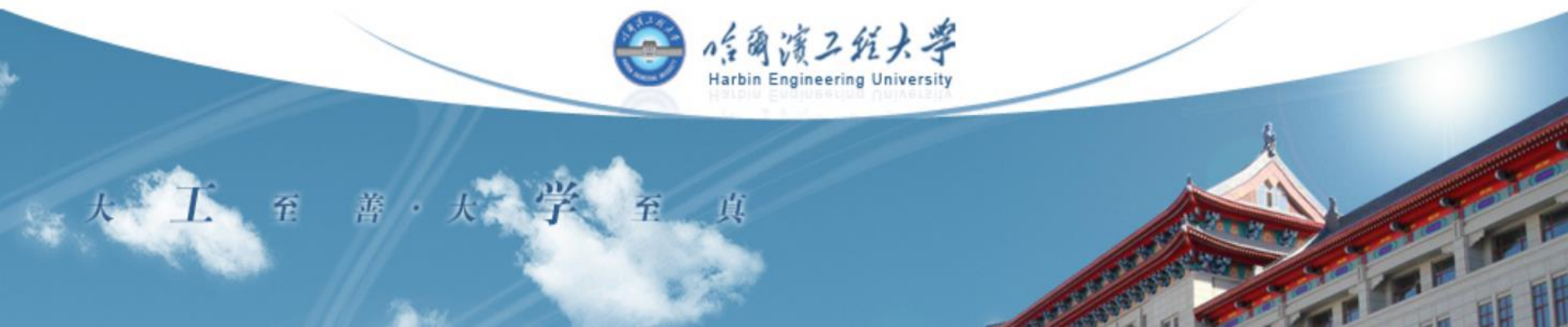
科技论文实验数据统计分析与可视化

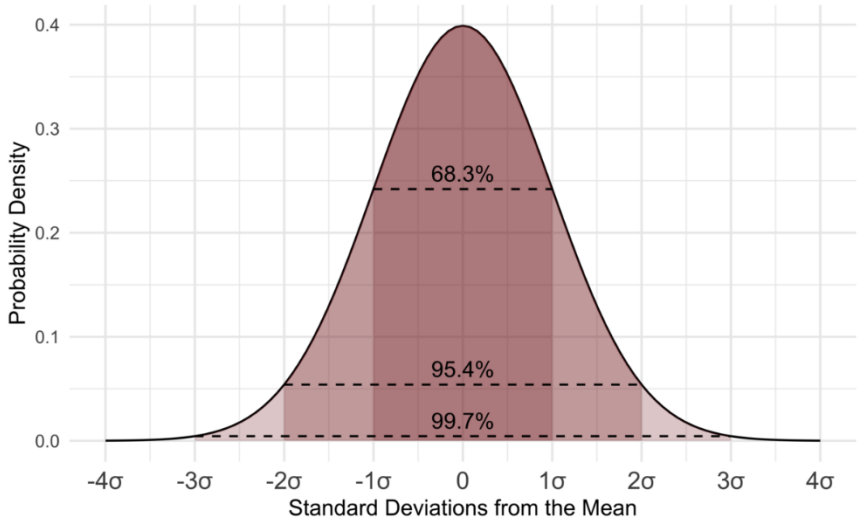
张立国

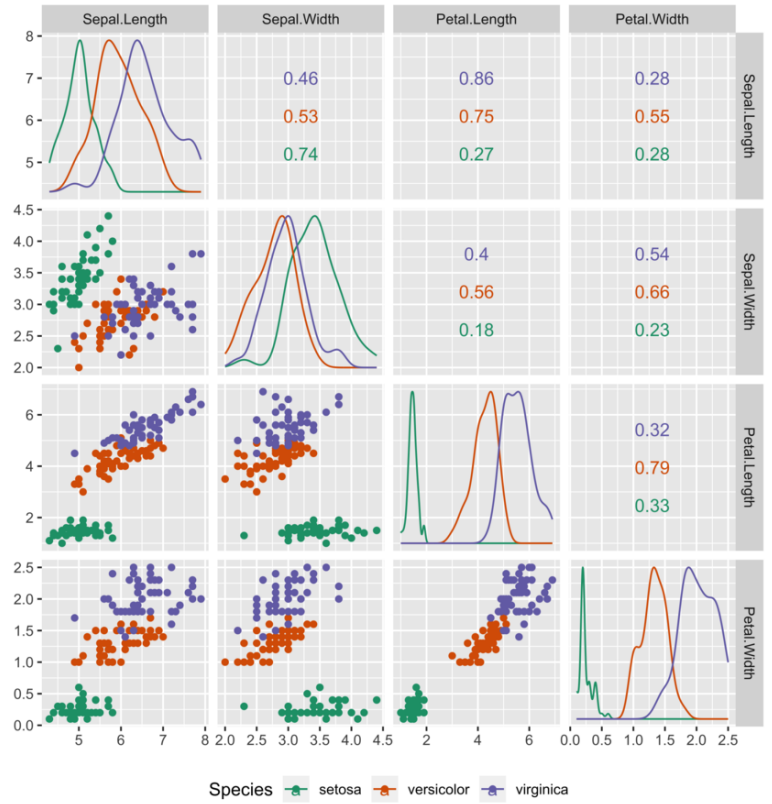
哈尔滨工程大学
Harbin Engineering University



数据统计、分析与可视化

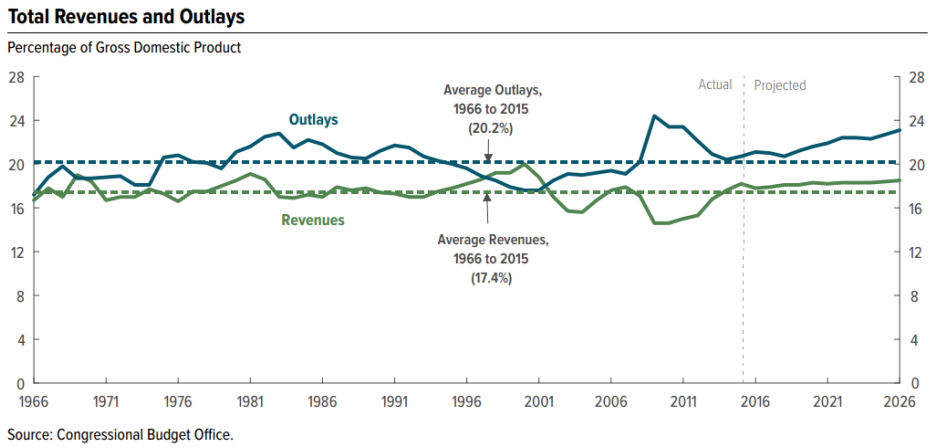
数据统计 是关于数据的收集，整理，分析，解释和展示的一项研究。在将统计数据应用于科学、工业或社会问题时，通常从统计数据或要研究的统计模型开始。统计学处理数据的领域包括根据调查和实验的设计进行数据收集的计划。



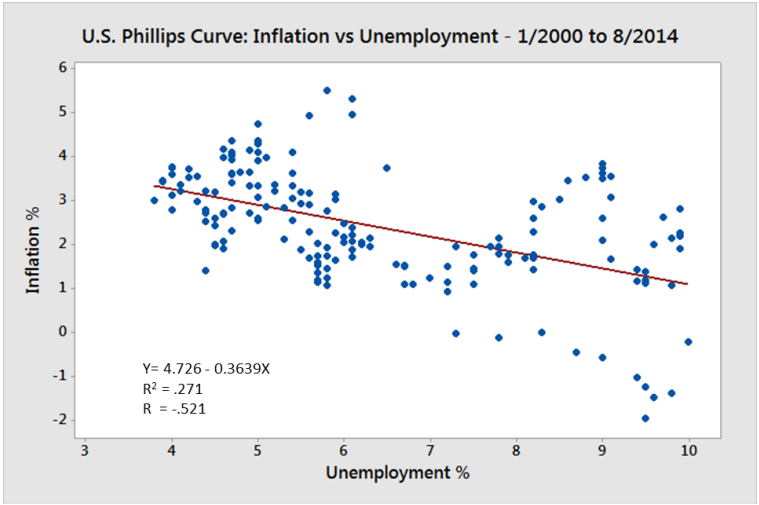


数据统计、分析与可视化

数据分析 是将原始数据进行排序和组织的过程，用于帮助解释过去和预测未来的一系列方法。数据分析不只是针对数字，而是关于如何设定或提出问题、演化解释，以及验证假说的过程。



用折线图说明的时间序列说明了美国联邦支出和收入随时间的趋势。



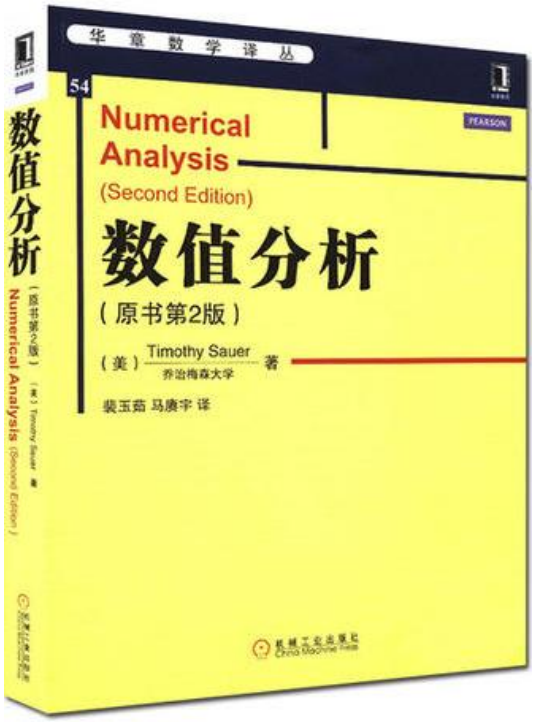
散点图，说明在各个时间点测得的两个变量（通货膨胀和失业）之间的相关性。

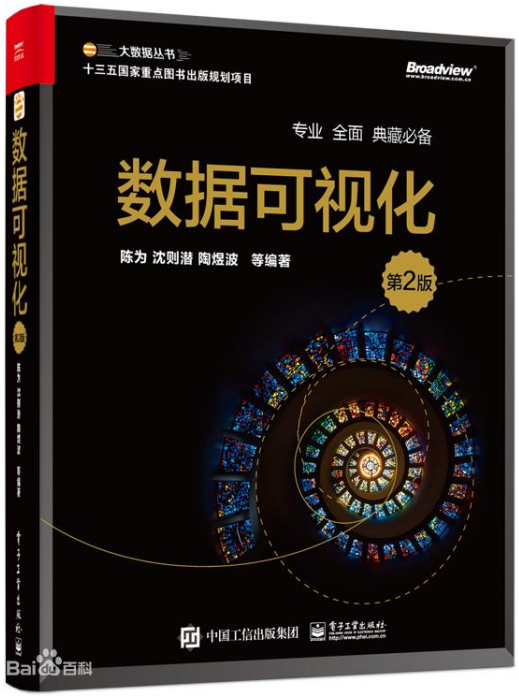
数据统计、分析与可视化

数据可视化 是利用人眼的感知能力对数据进行交互的可视表达，

数据统计、分析与可视化

参考书籍





数据统计

- 当无法收集普查数据时，统计学家会通过制定具体的实验设计和调查样本来收集数据。典型采样可确保推论和结论可以合理地从样本扩展到总体。
- 生成统计数据的测量过程也会出错。这些错误中有许多被分类为随机性（噪声）或系统性（固有性）（偏差），但也可能发生其他类型的错误（例如，失误，例如当分析人员报告不正确的单位时）。丢失数据或检查的存在可能导致估计偏差，并且已经开发出解决这些问题的特定技术。

数据统计

- 两种主要的统计方法：描述统计和推论统计。描述统计是指用平均数或标准差等指标从样本中总结数据;推论统计是指从服从随机变化影响的数据（如观测误差、样本方差）中得出结论。
- 标准的统计程序包括数据收集，检验两个数据集之间的关系，或一个数据集和从一个理想化的模型得出的合成数据之间的关系。对于两个数据集之间的统计关系提出了一个假设，并将其与理想化的两个数据集之间没有关系的零假设进行了比较。从零假设开始，可以识别出两种基本形式的错误:第一类错误(零假设被错误地拒绝，给出一个“假阳性”)和第二类错误(零假设没有被拒绝，群体之间的实际关系被忽略，给出一个“假阴性”)。

数据统计

接受者操作特性曲线（receiver operating characteristic curve，简称ROC曲线）

真阳性（TP）——等效 命中
真负　（TN）——等效 正确拒绝
误报　（FP）——等效 有误报，I型错误
假阴性（FN）——等效 错过，II型错误

预测 \ 实际		“金标准”结果		合计		
		阳性	阴性			
诊断结果	阳性	真阳性(TP)	假阳性(FP)	prediction positive = (TP+FP)	PPV = TP/prediction positive 又被称为正确率（Precision）	FDR = FP/prediction positive
	阴性	假阴性(FN)	真阴性(TN)	prediction negative = (FN+TN)	FOR = FN/prediction positive	NPV = TN/prediction positive
合计		condition positive = (TP+FN)	condition negative = (FP+TN)	N = TP+FN+FP+TN		
		真阳性率TPR = TP/condition positive 又被称为灵敏度（Sensitivity），召回率（Recall）	假阳性率FPR = FP/condition negative 又被称为误诊率=1-特异度			
		假阴性率FNR = FN/condition positive 又被称为漏诊率=1-灵敏度	真阴性率TNR = TN/condition negative 又被称为特异度（Specificity）			

数据统计

例：

灵敏度，召回率，命中率或真实阳性率（TPR）

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

特异性，选择性或真阴性率（TNR）

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = 1 - \mathrm{FPR}$$

精确度或阳性预测值（PPV）

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = 1 - \mathrm{FDR}$$

阴性预测值（NPV）

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} = 1 - \mathrm{FOR}$$

遗漏率或误报率（FNR）

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{P}} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} = 1 - \mathrm{TPR}$$

脱落或假阳性率（FPR）

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR}$$

错误发现率（FDR）

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}} = 1 - \mathrm{PPV}$$

错误遗漏率（FOR）

$$\mathrm{FOR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TN}} = 1 - \mathrm{NPV}$$

患病阈值（PT）

$$PT = \frac{\sqrt{TPR(-TNR+1)} + TNR - 1}{(TPR + TNR - 1)}$$

威胁评分（TS）或关键成功指数（CSI）

$$\mathrm{TS} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}$$

准确度（ACC）

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

平衡精度（BA）

$$\mathrm{BA} = \frac{TPR + TNR}{2}$$

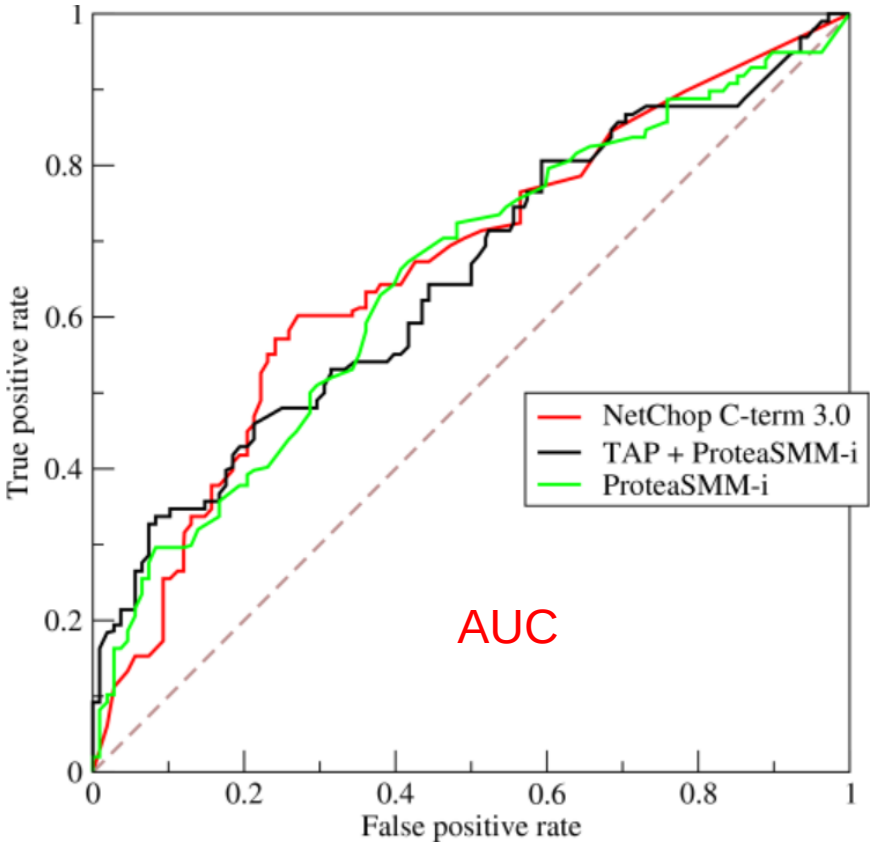
数据统计

例：

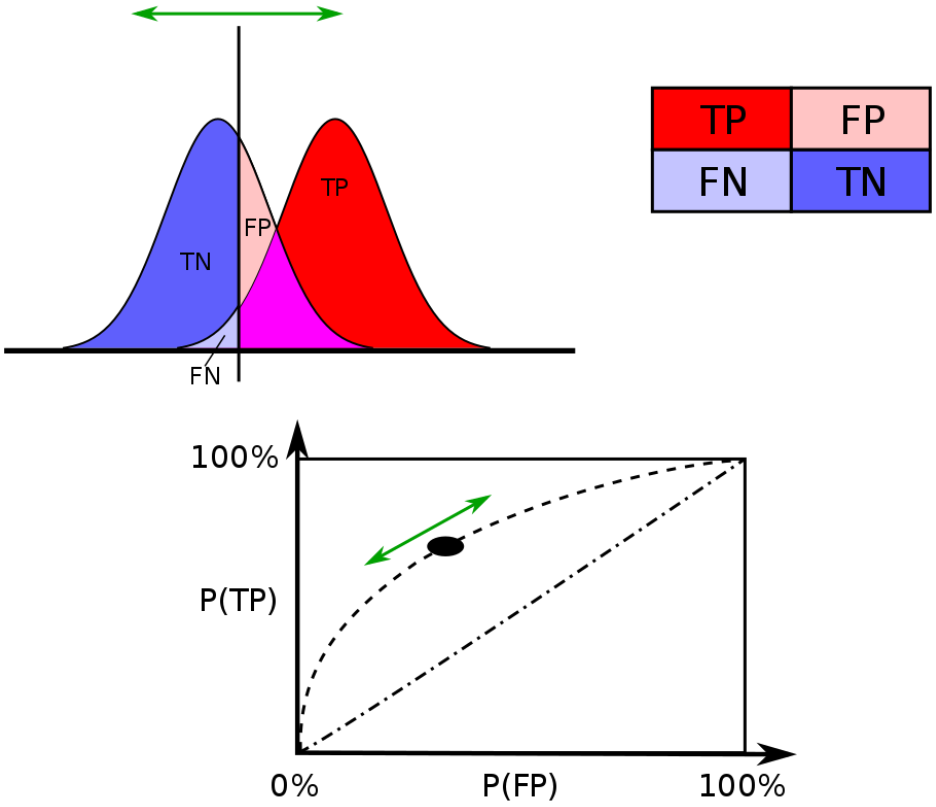
A	B	C	C′
TP=63 FP=28 91 FN=37 TN=72 109 100 100 200	TP=77 FP=77 154 FN=23 TN=23 46 100 100 200	TP=24 FP=88 112 FN=76 TN=12 88 100 100 200	TP=76 FP=12 88 FN=24 TN=88 112 100 100 200
TPR = 0.63	TPR = 0.77	TPR = 0.24	TPR = 0.76
FPR = 0.28	FPR = 0.77	FPR = 0.88	FPR = 0.12
PPV = 0.69	PPV = 0.50	PPV = 0.21	PPV = 0.86
F1 = 0.66	F1 = 0.61	F1 = 0.23	F1 = 0.81
ACC = 0.68	ACC = 0.50	ACC = 0.18	ACC = 0.82

数据统计

例：



ROC曲线的三种预测肽。裂解蛋白酶体



病人和健康人的血液蛋白质成分分布，及其患病ROC曲线

数据统计

混淆矩阵

混淆矩阵是ROC曲线绘制的基础，同时它也是衡量分类型模型准确度中最基本，最直观，计算最简单的方法。

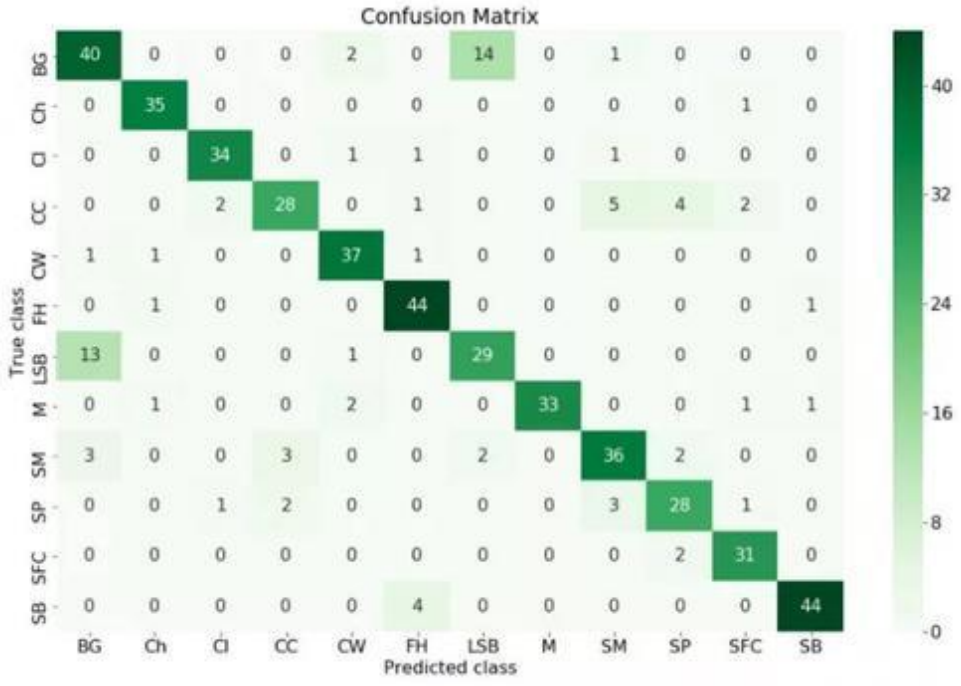
		Actual class	
		Cat	Dog
Predicted class	Cat	5	2
	Dog	3	3

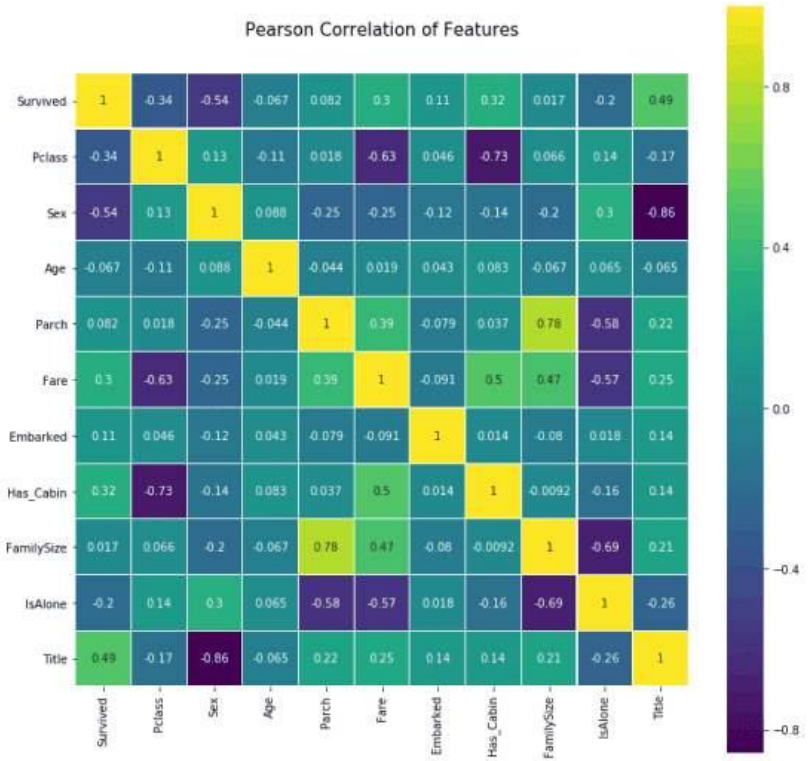
		Actual class	
		Cat	Non-cat
Predicted class	Cat	5 True Positives	2 False Positives
	Non-cat	3 False Negatives	3 True Negatives

数据统计

混淆矩阵

混淆矩阵是ROC曲线绘制的基础，同时它也是衡量分类型模型准确度中最基本，最直观，计算最简单的方法。

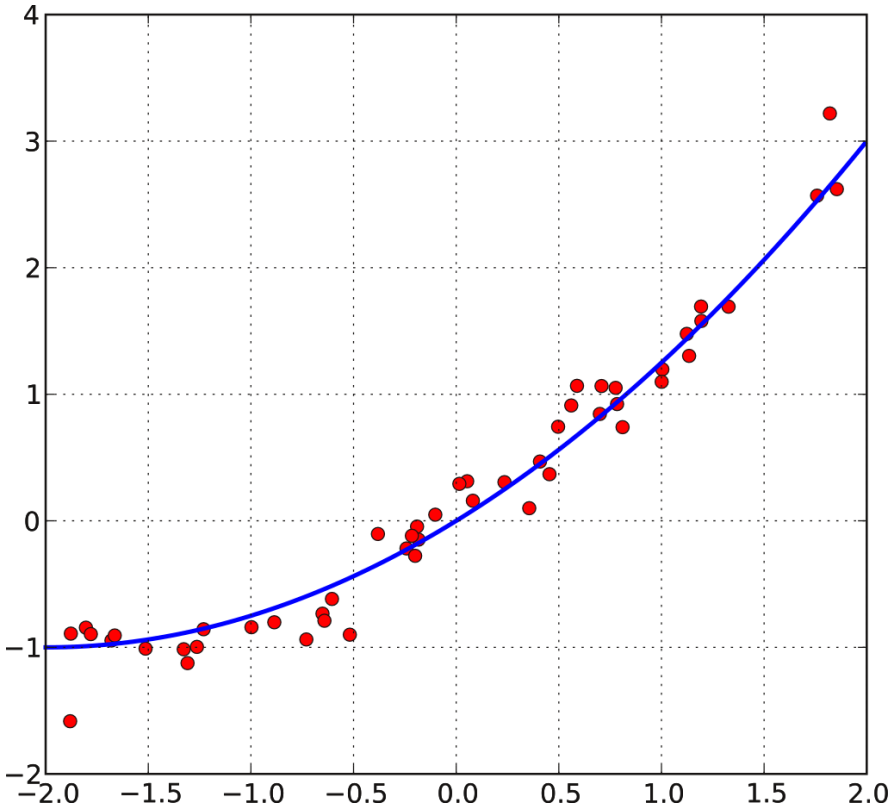


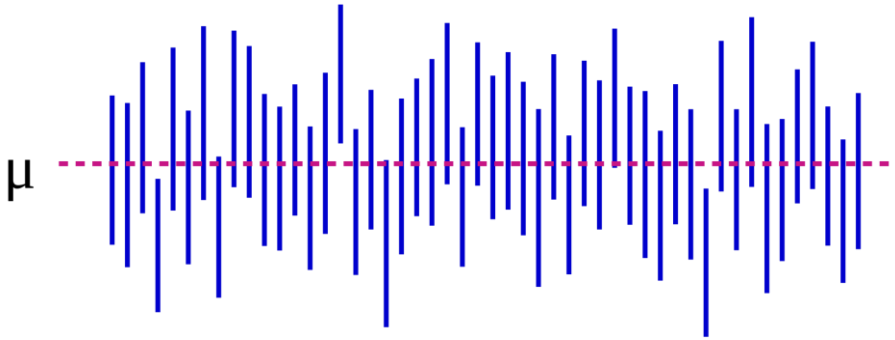


数据统计

误差 errors

误差是测量测得的量值减去参考量（真实值）的结果。





误差图

均方误差 $MSE = \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - y_i)^2$

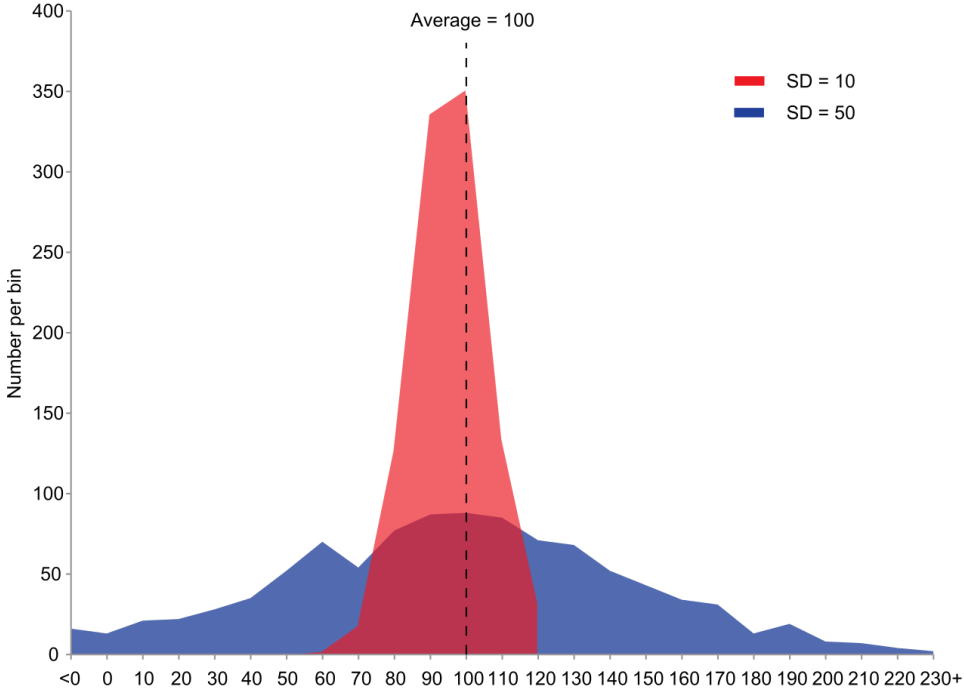
红色为点为测量值，蓝色曲线为真实值

数据统计

方差

方差是在概率论和统计方差衡量随机变量或一组数据时离散程度的度量。概率论中方差用来度量随机变量和其数学期望（即均值）之间的偏离程度。

$$
\begin{aligned}
\operatorname{Var}(X) &= \operatorname{E}\left[(X - \operatorname{E}[X])^2\right] \\
&= \operatorname{E}\left[X^2 - 2X\operatorname{E}[X] + \operatorname{E}[X]^2\right] \\
&= \operatorname{E}\left[X^2\right] - 2\operatorname{E}[X]\operatorname{E}[X] + \operatorname{E}[X]^2 \\
&= \operatorname{E}\left[X^2\right] - \operatorname{E}[X]^2
\end{aligned}
$$



来自两组集合的均值相同但方差不同的样本示例。红色的总体具有均值100和方差100，而蓝色的总体具有均值100和方差2500。

数据统计

标准差

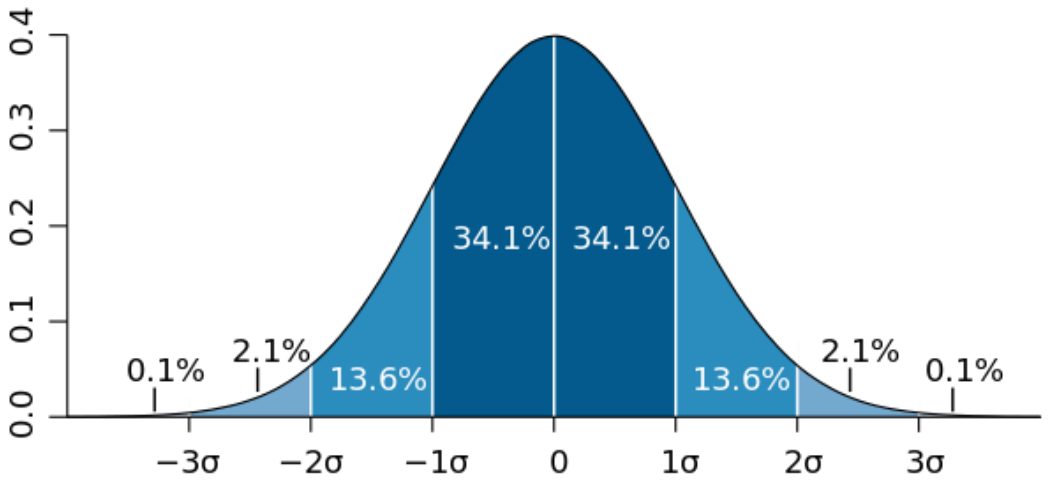
标准差（Standard Deviation），是离均差平方的算术平均数的算术平方根，用σ表示。标准差也被称为标准偏差，或者实验标准差，在概率统计中最常使用作为统计分布程度上的测量依据。

总体标准差：$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$$

样本标准差：$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

标准误差：$$\sigma_n = \frac{\sigma}{\sqrt{n}}$$

与方差：方差=标准差的平方。

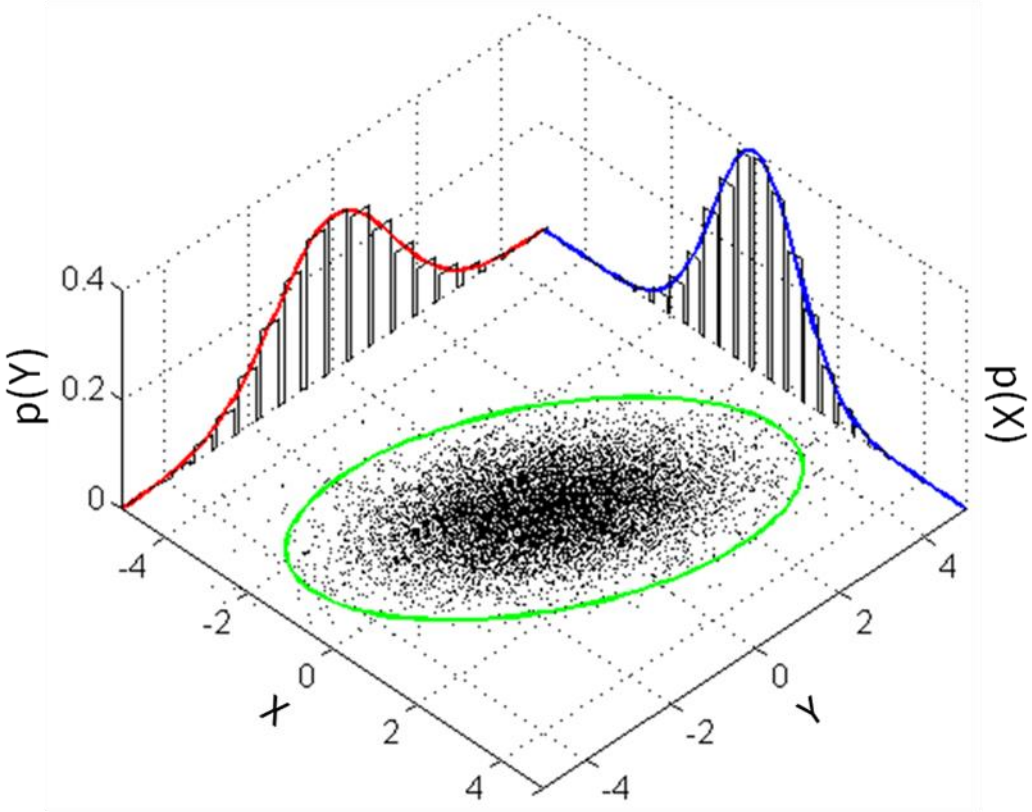


深蓝色区域是距均值小于一个标准差之内的数值范围，在高斯分布中，该范围占比为全部数值的68.2%。两个标准差之内占比95.4%，三个标准差之内占比为99.6%。

数据统计

标准差

标准差（Standard Deviation），是离均差平方的算术平均数的算术平方根，用σ表示。标准差也被称为标准偏差，或者实验标准差，在概率统计中最常使用作为统计分布程度上的测量依据。

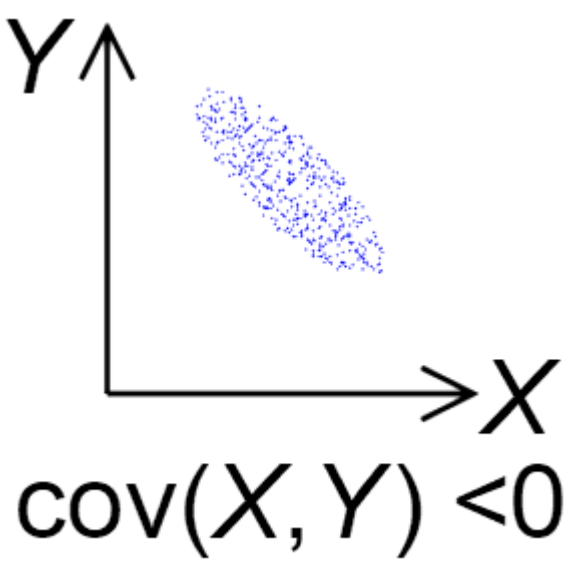


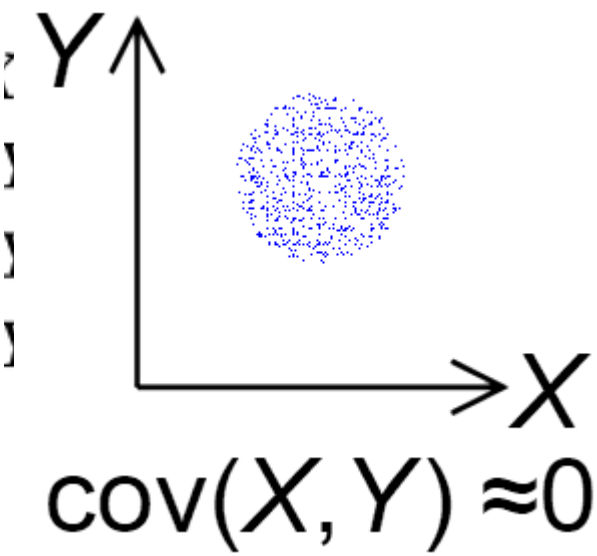
二维正态分布的标准偏差椭圆（绿色）。

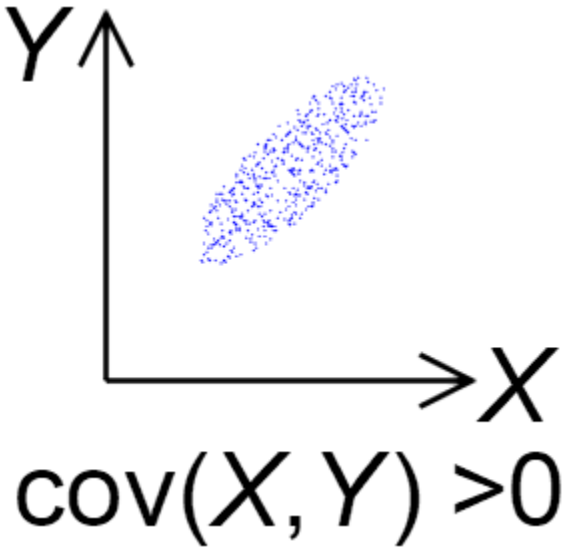
数据统计

协方差（Covariance）

协方差表示的是两个变量的总体的误差。 如果两个变量的变化趋势一致，也就是说如果其中一个大于自身的期望值，另外一个也大于自身的期望值，那么两个变量之间的协方差就是正值。如果两个变量的变化趋势相反，即其中一个大于自身的期望值，另外一个却小于自身的期望值，那么两个变量之间的协方差就是负值。





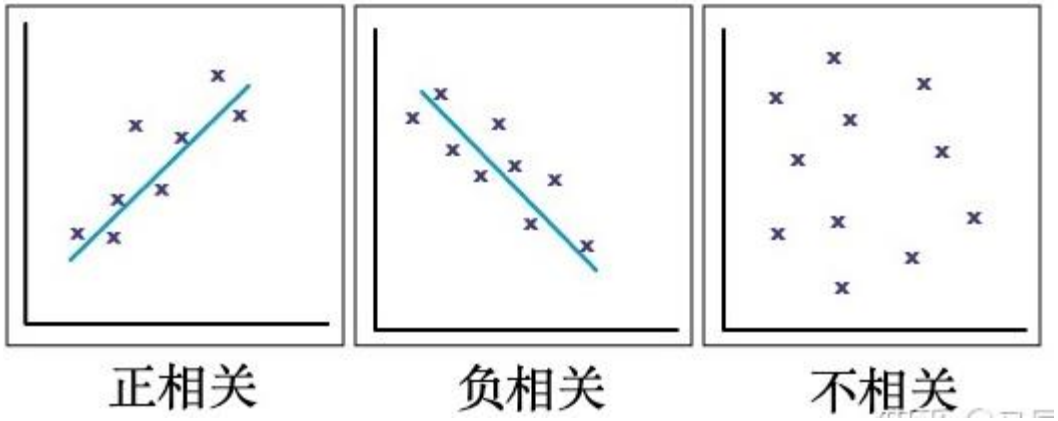


数据统计

相关系数（Correlation coefficient）

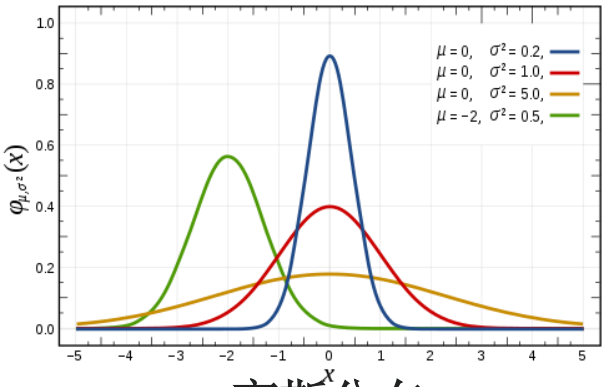
相关系数是协发差的归一化(normalization)， 消除了两个变量量纲/变化幅度不同的影响。单纯反映两个变量在每单位变化的相似程度。

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

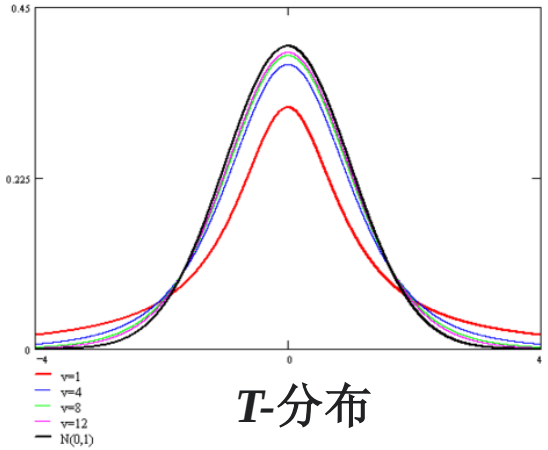


数据统计

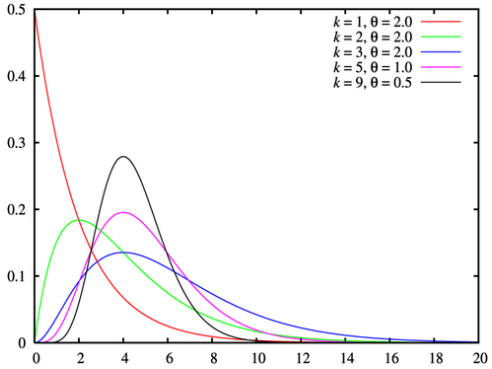
常见概率分布



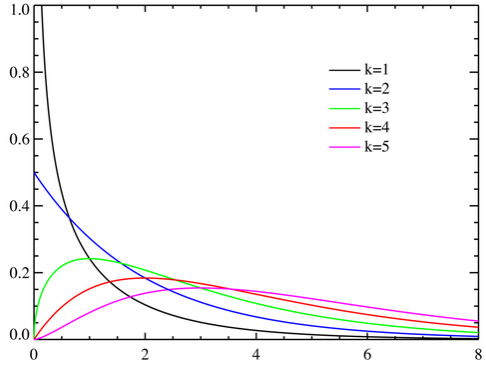
高斯分布



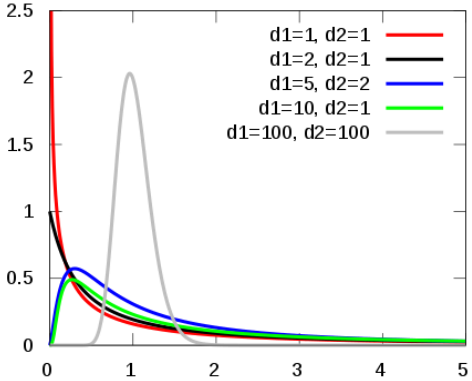
T-分布



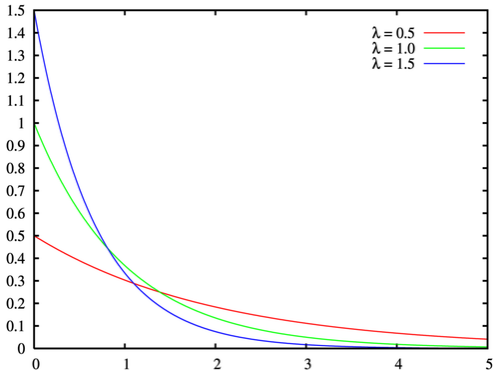
Γ-分布



χ^2分布



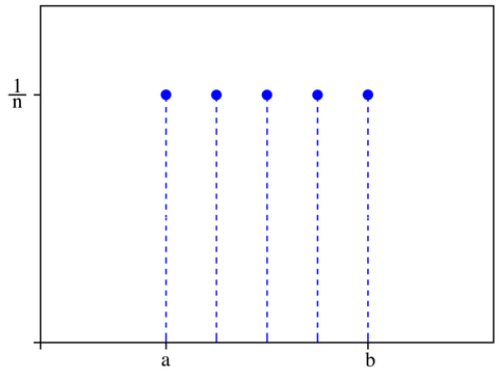
F-分布



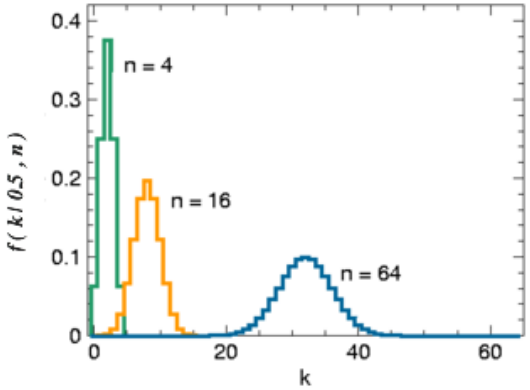
指数分布

数据统计

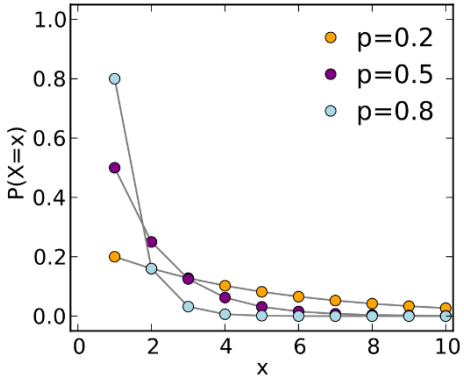
常见概率分布



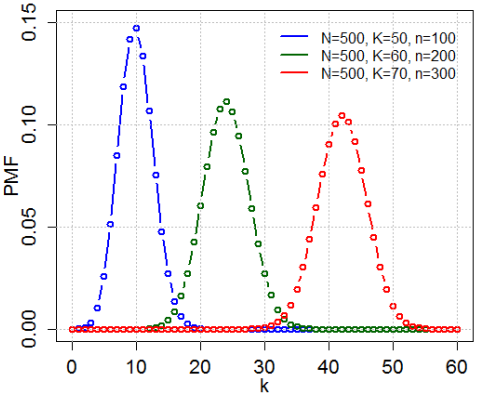
均匀分布



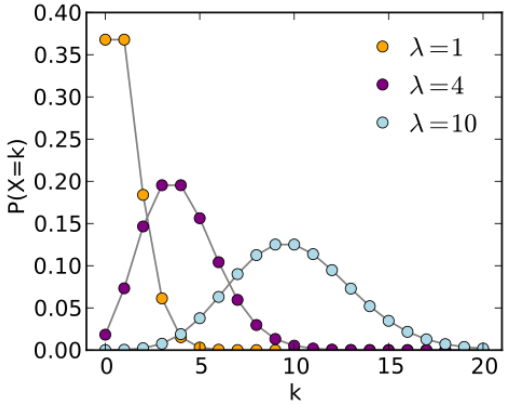
二项分布



几何分布



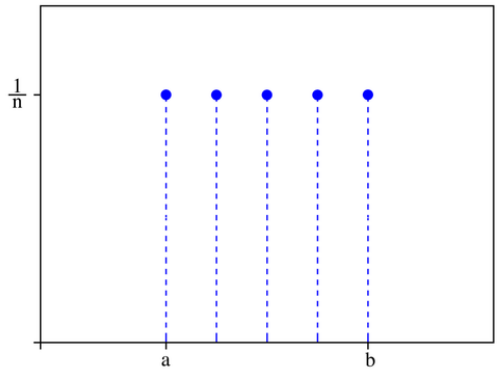
超几何分布



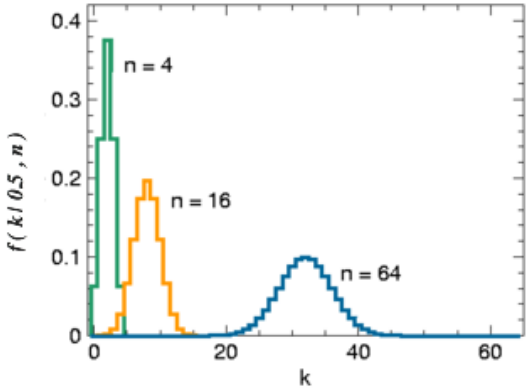
泊松分布

数据统计

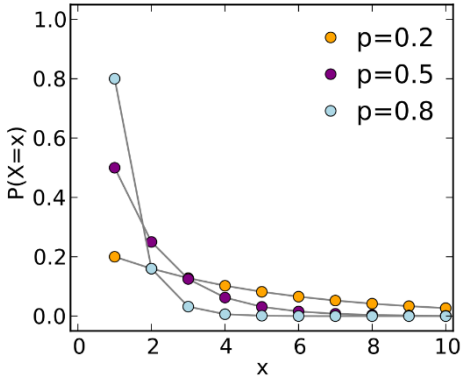
常见概率分布



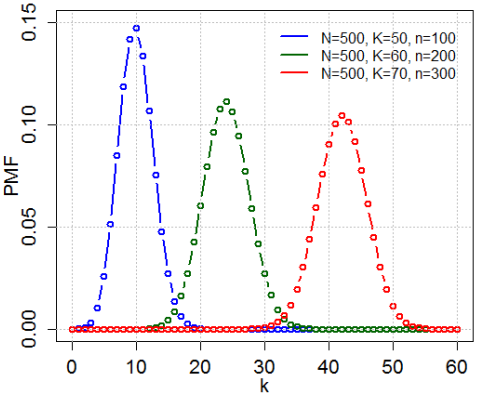
均匀分布



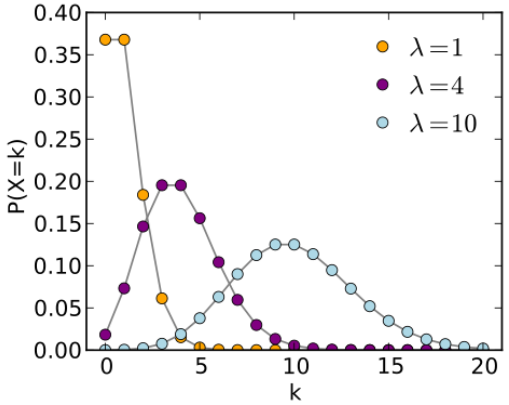
二项分布



几何分布



超几何分布



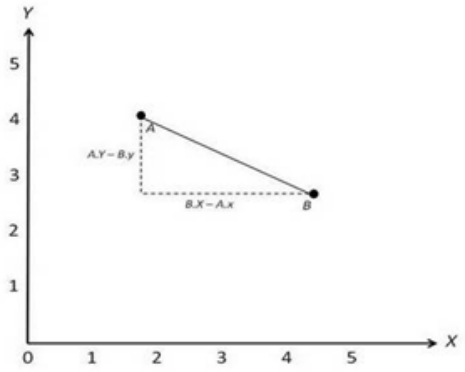
泊松分布

数据统计

常见相似度计算

欧几里德距离

$$d(x,y)=\sqrt{(\sum(x_i-y_i)^2)}$$



余弦相似度

$$sim(X,Y)=cos\theta=\frac{\vec{x}\cdot\vec{y}}{||x||\cdot||y||}$$