# 统计学方法及其应用

# Statistical Methods with Applications

**Rui Jiang, PhD**

**Associate Professor**

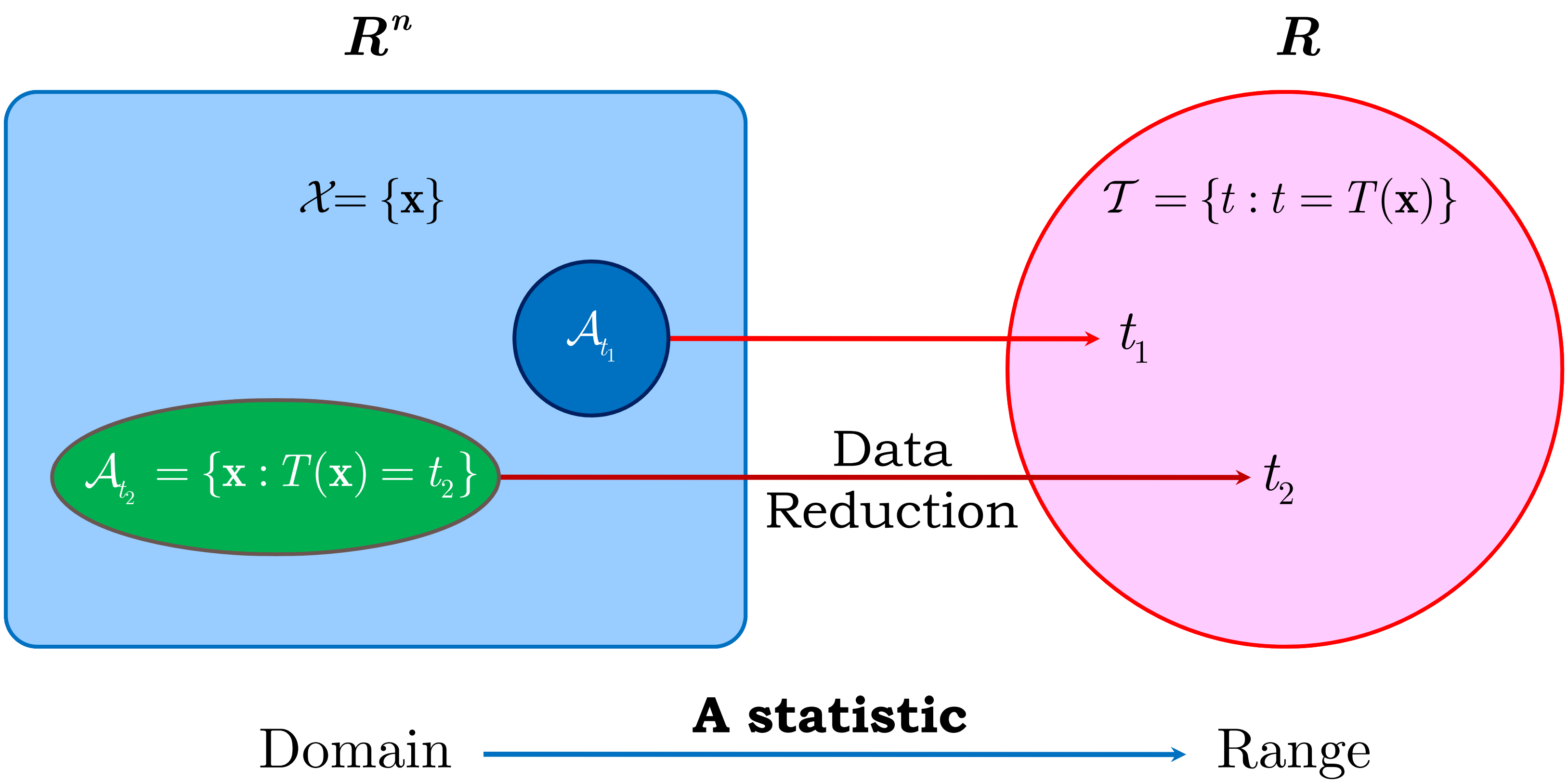
Ministry of Education Key Laboratory of Bioinformatics
Bioinformatics Division, TNLIST/Department of Automation
Tsinghua University, Beijing 100084, China

# Key points of statistics

- Population
  - A distribution that we are unable to see but interested in
- Sample
  - A set of iid random variables sampled from the population
- Statistic
  - Summary of the sample, reduction of the data
  - Identical observations of samples lead to equal values of statistics
  - Equal values of statistics do not mean identical observations of samples

$$\begin{aligned} \mathbf{x} = \mathbf{y} \quad &\Rightarrow \quad T(\mathbf{x}) = T(\mathbf{y}) \\ T(\mathbf{x}) = T(\mathbf{y}) \quad &\not\Rightarrow \quad \mathbf{x} = \mathbf{y} \end{aligned}$$

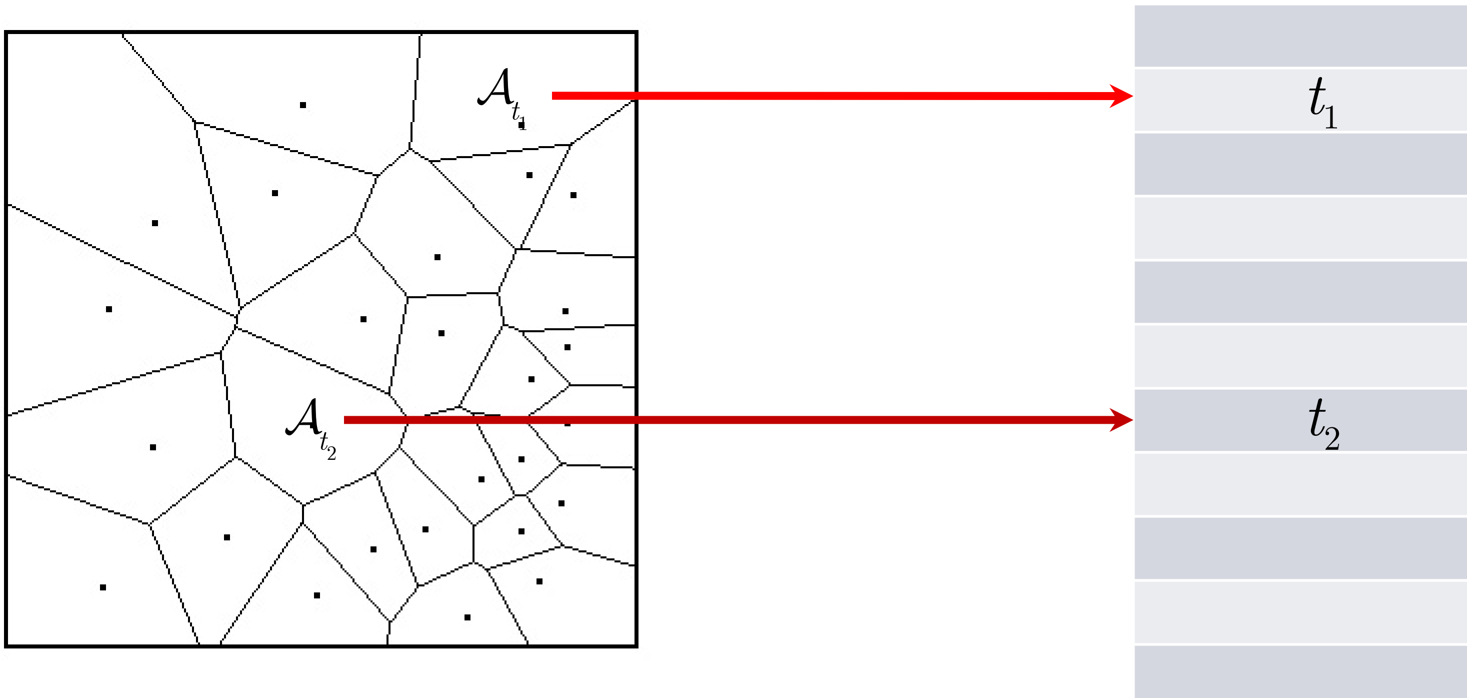
# Data reduction




Report a small number of data instead of a large number of data

# Sample space partition

$\mathcal{X} = \{\mathbf{x}\}$ $\qquad\qquad$ $\mathcal{T} = \{t : t = T(\mathbf{x})\}$




$$\text{Domain} \xrightarrow{\textbf{A statistic}} \text{Range}$$

$$\mathcal{A}_{t_i} = \{\mathbf{x} : T(\mathbf{x}) = t_i\}$$

A statistic implies a partition of the sample space

# Principles of Data Reduction

统计学方法及其应用 统计学基础 数据简约的原理

*"A random variable is a quantity whose values are random and to which a probability distribution is assigned."*

# The sufficiency principle

## SUFFICIENCY PRINCIPLE

If $T(\mathbf{X})$ is a ***sufficient statistic*** for $\theta$, then any inference about $\theta$ should depend on the sample $\mathbf{X}$ only through the value $T(\mathbf{X})$. That is, if $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that $T(\mathbf{x}) = T(\mathbf{y})$, then the inference about $\theta$ should be the same whether $\mathbf{X}=\mathbf{x}$ or $\mathbf{X}=\mathbf{y}$ is observed.

A sufficient statistic captures **ALL** the information about the parameter contained in the sample. Any additional information in the sample, besides the value of the sufficient statistic, does **not** contain any more information about the parameter.

# Sufficient statistics

## *Sufficient statistics*

A statistic $T(\mathbf{X})$ is a sufficient statistic for $\theta$ if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on $\theta$.

$$P_\theta(\mathbf{X}=\mathbf{x} \mid T(\mathbf{X})=T(\mathbf{x})) = \frac{P_\theta(\mathbf{X}=\mathbf{x} \text{ and } T(\mathbf{X})=T(\mathbf{x}))}{P_\theta(T(\mathbf{X})=T(\mathbf{x}))}$$

$$= \frac{P_\theta(\mathbf{X}=\mathbf{x})}{P_\theta(T(\mathbf{X})=T(\mathbf{x}))}$$

$$P(\mathbf{X}=\mathbf{x} \mid T(\mathbf{X})=T(\mathbf{x})) = \frac{p(\mathbf{x} \mid \theta)}{q(T(\mathbf{x}) \mid \theta)}$$

# Sufficient statistics

***Sufficient condition***

If $p(\mathbf{x} \mid \theta)$ is the joint pdf or pmf of $\mathbf{X}$ and $q(t \mid \theta)$ is the pdf or pmf of $T(\mathbf{X})$, then $T(\mathbf{X})$ is a sufficient statistic for $\theta$ if, for every $\mathbf{x}$ in the sample space, the ratio

$$\frac{p(\mathbf{x} \mid \theta)}{q(T(\mathbf{x}) \mid \theta)}$$

is constant as a function of $\theta$.

# Binomial sufficient statistic

Let $X_1, \ldots, X_n$ be iid Bernoulli random variables with parameter $\theta$, where $0 < \theta < 1$. Define the statistic $T(\mathbf{X}) = X_1 + \cdots + X_n = \sum_{i=1}^{n} X_i$. Then,

$$p(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n-\sum_{i=1}^{n} x_i} = \theta^{t} (1-\theta)^{n-t}$$

$$q(T(\mathbf{x}) \mid \theta) = \binom{n}{t} \theta^{t} (1-\theta)^{1-t}$$

$$\frac{p(\mathbf{x} \mid \theta)}{q(T(\mathbf{x}) \mid \theta)} = \frac{\theta^{t} (1-\theta)^{n-t}}{\binom{n}{t} \theta^{t} (1-\theta)^{1-t}} = \binom{n}{t}^{-1} = \binom{n}{\sum_{i=1}^{n} x_i}^{-1}$$

**The total number of successes in a Bernoulli sample is a sufficient statistic for the ratio of success.**

# Normal sufficient statistic

Let $X_1,\ldots,X_n$ be iid random variables with common pdf $N(\mu,\sigma^2)$, where $\sigma^2$ is known. Define the statistic $T(\mathbf{X}) = \bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$. Then,

$$p(\mathbf{x} \mid \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2\right]$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \exp\left[-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right]$$

$$q(T(\mathbf{x}) \mid \mu) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left[-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right]$$

$$\frac{p(\mathbf{x} \mid \mu)}{q(T(\mathbf{x}) \mid \mu)} = n^{-\frac{1}{2}}(2\pi\sigma^2)^{-\frac{n-1}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

**The sample mean is a sufficient statistic for the population mean when population variance is known.**

# Sufficient order statistics

Let $X_1, \ldots, X_n$ be iid random variables from a certain pdf $f(x)$, about which we are unable to specific any more information. Define the statistic $T(\mathbf{X}) = (X_{(1)}, \ldots, X_{(n)})$.

Then,

$$q(T(\mathbf{x})) = f_{X_{(1)}, \ldots, X_{(n)}}(x_1, \ldots, x_n) = n!\, f_X(x_1) f_X(x_2) \cdots f_X(x_n) \propto p(\mathbf{x})$$

**The vector of all order statistics is a sufficient statistic for the unknown population $f(x)$.**

Outside the exponential family, do not waste your time, just use order statistics

# Factorization theorem

***Sufficient and necessary condition***

Let $f(\mathbf{x} \mid \theta)$ denote the joint pdf or pmf of a sample $\mathbf{X}$.

A statistic $T(\mathbf{X})$ is a sufficient statistic for $\theta$ if and only if there exist functions $g(t \mid \theta)$ and $h(\mathbf{x})$ such that, for all sample points $\mathbf{x}$ and all parameter points $\theta$,

$$f(\mathbf{x} \mid \theta) = g(T(\mathbf{x}) \mid \theta)h(\mathbf{x}).$$

# Sufficiency

If there exist functions $g(t \mid \theta)$ and $h(\mathbf{x})$ such that

$$f(\mathbf{x} \mid \theta) = g(T(\mathbf{x}) \mid \theta) h(\mathbf{x}),$$

then $T(\mathbf{X})$ is a sufficient statistic for $\theta$.

Let $q(t \mid \theta)$ be the pmf of $T(\mathbf{X})$, examine the ratio $f(\mathbf{x} \mid \theta) \,/\, q(T(\mathbf{x}) \mid \theta)$.

$$\begin{aligned}
\frac{f(\mathbf{x} \mid \theta)}{q(T(\mathbf{x}) \mid \theta)} &= \frac{g(T(\mathbf{x}) \mid \theta) h(\mathbf{x})}{q(T(\mathbf{x}) \mid \theta)} \\
&= \frac{g(T(\mathbf{x}) \mid \theta) h(\mathbf{x})}{\sum_{A_{T(\mathbf{x})}} g(T(\mathbf{y}) \mid \theta) h(\mathbf{y})} \quad A_{T(\mathbf{x})} = \{\mathbf{y} : T(\mathbf{y}) = T(\mathbf{x})\} \\
&= \frac{g(T(\mathbf{x}) \mid \theta) h(\mathbf{x})}{g(T(\mathbf{x}) \mid \theta) \sum_{A_{T(\mathbf{x})}} h(\mathbf{y})} \\
&= \frac{h(\mathbf{x})}{\sum_{A_{T(\mathbf{x})}} h(\mathbf{y})}
\end{aligned}$$

Independent of $\theta$, therefore, $T(\mathbf{X})$ is a sufficient statistic for $\theta$.

# Necessity

If $T(\mathbf{X})$ is a sufficient statistic for $\theta$, then there exist functions $g(t \mid \theta)$ and $h(\mathbf{x})$ such that

$$f(\mathbf{x} \mid \theta) = g(T(\mathbf{x}) \mid \theta)h(\mathbf{x})$$

Choose

$$\begin{aligned} g(T(\mathbf{x}) \mid \theta) &= P_\theta(T(\mathbf{X}) = T(\mathbf{x})), \text{ the pmf of } T(\mathbf{X}) \\ h(\mathbf{x}) &= P(\mathbf{X} = x \mid T(\mathbf{X}) = T(\mathbf{x})) \end{aligned}$$

Since $T(\mathbf{X})$ is a sufficient statistic for $\theta$, $h(\mathbf{x})$ does not depend on $\theta$. We now show that the product of the above valid choice yields the pmf of $\mathbf{X}$.

$$\begin{aligned} f(\mathbf{x} \mid \theta) &= P_\theta(\mathbf{X} = \mathbf{x}) \\ &= P_\theta(\mathbf{X} = \mathbf{x} \textbf{ AND } T(\mathbf{X}) = T(\mathbf{x})) \\ &= P_\theta(T(\mathbf{X}) = T(\mathbf{x}))P_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})) \\ &= P_\theta(T(\mathbf{X}) = T(\mathbf{x}))P(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})) \\ &= g(T(\mathbf{x}) \mid \theta)h(\mathbf{x}) \end{aligned}$$

# Normal sufficient statistic

Let $X_1,\ldots,X_n$ be iid random variables with common pdf $N(\mu,\sigma^2)$. Define statistics

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \text{ and } S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

Then,

$$\begin{aligned}
f(\mathbf{x}\mid\mu,\sigma^2) &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]\\
&=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2\right]\\
&=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(\bar{x}-\mu)^2+\underbrace{2\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-\mu)}_{=0}\right]\right\}\\
&=\underbrace{(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2-\frac{n-1}{2\sigma^2}s^2\right]}_{g\left(\bar{x},s^2\mid\mu,\sigma^2\right)}\times\underbrace{1}_{h(\mathbf{x})}
\end{aligned}$$

**The vector of the sample mean and the sample variance is a sufficient statistic in the case that the population variance is unknown.**

# Exponential family

**_Sufficient statistic for the exponential family_**

Let $X_1, \ldots, X_n$ be iid random variables from a pdf or pmf that belongs to an exponential family given by

$$f(x \mid \boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left[\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right],$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d), d \leq k$. Then,

$$T(X) = \left(\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is a sufficient statistic for $\boldsymbol{\theta}$.

# Normal sufficient statistics

Normal pdf

$$\varphi(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Written it as exponential family,

$$\varphi(x \mid \mu, \sigma^2) = \underbrace{1}_{h(x)} \cdot \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right)}_{c(\mu,\sigma^2)} \exp\left[\underbrace{\frac{\mu}{\sigma^2}}_{w_1(\mu,\sigma^2)} \underbrace{x}_{t_1(x)} + \left(\underbrace{-\frac{1}{2\sigma^2}}_{w_2(\mu,\sigma^2)} \underbrace{x^2}_{t_2(x)}\right)\right].$$

Thus for a sample $X_1, \ldots, X_n$, a sufficient statistic for $(\mu, \sigma^2)$ is

$$\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right).$$

# Sufficient statistic is not unique

Let $X_1,\ldots,X_n$ be iid random variables with common pdf $N(\mu,\sigma^2)$, where $\sigma^2$ is known. Define

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Then,

$$\begin{aligned}
f(\mathbf{x} \mid \mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2\right] \\
&= \underbrace{(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2 - \frac{n-1}{2\sigma^2}s^2\right]}_{g(\bar{x},s^2|\mu)} \times \underbrace{1}_{h(\mathbf{x})} \\
&= \underbrace{\exp\left[-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right]}_{g(\bar{x}|\mu)} \underbrace{(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]}_{h(x)}
\end{aligned}$$

# Two trivial sufficient statistics

Let $X_1,\ldots,X_n$ be iid random variables from a certain pdf $f(x\mid\theta)$.
Define the statistic $T(\mathbf{X})=(X_1,\ldots,X_n)$, then,

$$f(\mathbf{x}\mid\theta)=\prod_{i=1}^{n}f(x_i\mid\theta)=\underbrace{\prod_{i=1}^{n}f(x_i\mid\theta)}_{g(T(x)|\theta)}\times\underbrace{1}_{h(x)}$$

Define the statistic $T(\mathbf{X})=(X_{(1)},\ldots,X_{(n)})$, then,

$$f(\mathbf{x}\mid\theta)=\prod_{i=1}^{n}f(x_i\mid\theta)=\underbrace{\prod_{i=1}^{n}f(x_{(i)}\mid\theta)}_{g(T(x)|\theta)}\times\underbrace{1}_{h(x)}$$

**The complete sample is always a sufficient statistic!**
**The vector of all order statistics is always a sufficient statistic!**

# Functions of a sufficient statistic

Suppose $T(\mathbf{X})$ is a sufficient statistic, by the Factorization Theorem, there exist $g$ and $h$ such that

$$f(\mathbf{x} \mid \theta) = g(T(\mathbf{x}) \mid \theta)h(\mathbf{x}).$$

Now, define $T^\star(\mathbf{x}) = r(T(\mathbf{x}))$ for all $\mathbf{x}$, where $r$ is a one-to-one function with inverse $r^{-1}$. Then,

$$f(\mathbf{x} \mid \theta) = g(T(\mathbf{x}) \mid \theta)h(\mathbf{x}) = g(r^{-1}(T^\star(\mathbf{x})) \mid \theta)h(\mathbf{x}).$$

Define a new function $g^\star(t \mid \theta) = g(r^{-1}(t) \mid \theta)$, we see that

$$f(\mathbf{x} \mid \theta) = g^\star(T^\star(\mathbf{x}) \mid \theta)h(\mathbf{x}).$$

Again, by the Factorization Theorem, $T^\star(\mathbf{x})$ is a sufficient statistic.

**Any one-to-one function of a sufficient statistic is a sufficient statistic**

# Minimal sufficient statistics

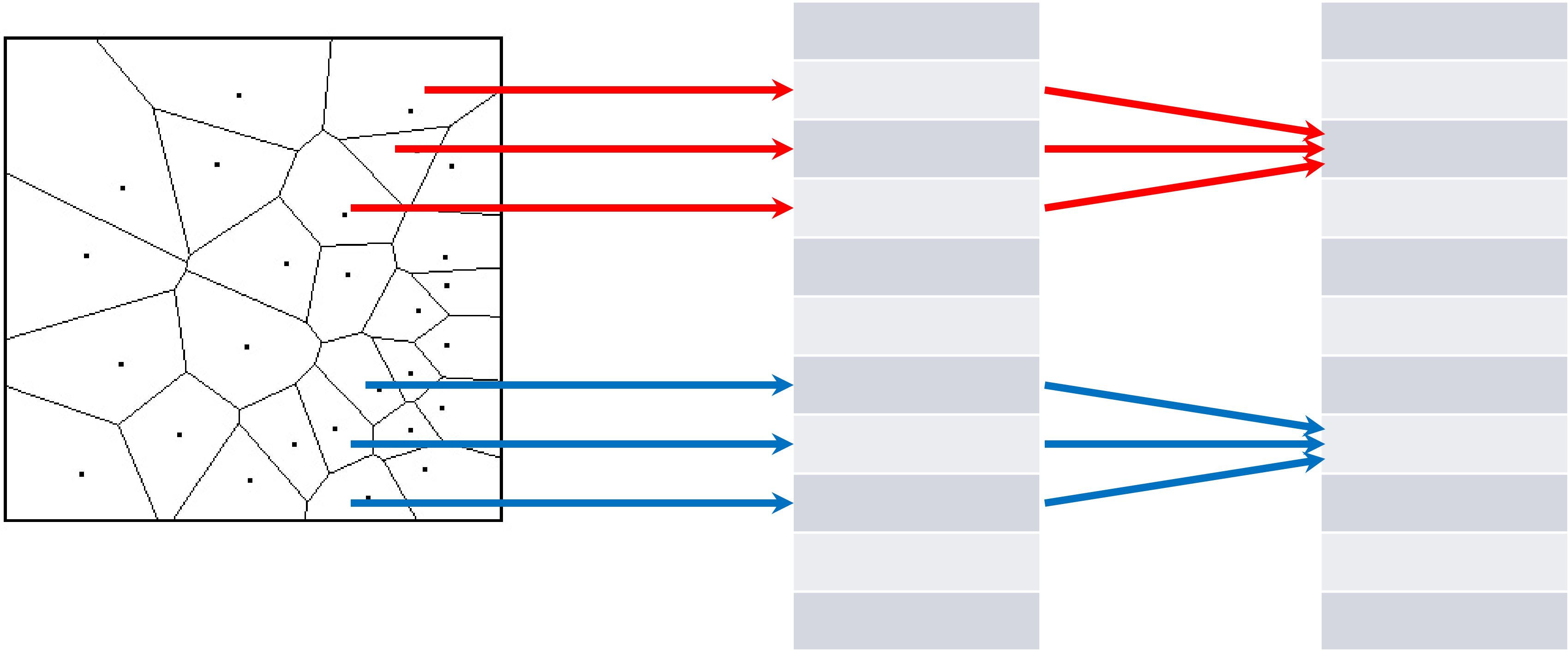
***Minimal sufficient statistics***

A sufficient statistic $T(\mathbf{X})$ is called a minimal sufficient statistic if, for any other sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{X})$ is a function of $T'(\mathbf{X})$. Or simply,

$$\text{if } T'(\mathbf{x}) = T'(\mathbf{y}), \text{ then } T(\mathbf{x}) = T(\mathbf{y}).$$

# Coarsest partition of the sample space

$\mathcal{X} = \{\mathbf{x}\}$ $\mathcal{T}' = \{t : t = T'(\mathbf{x})\}$ $\mathcal{T} = \{t : t = T(\mathbf{x})\}$



# Normal minimal sufficient statistic

Let $X_1,\dots,X_n$ be iid random variables with common pdf $N(\mu,\sigma^2)$, where $\sigma^2$ is known.

Define

$$T(\mathbf{X}) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } S(\mathbf{X}) = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Then, both

$$T_1(\mathbf{X}) = T(\mathbf{X}) = \bar{X}$$

and

$$T_2(\mathbf{X}) = (T(\mathbf{X}), S(\mathbf{X})) = (\bar{X}, S^2)$$

are sufficient statistics of the population mean.

However, if we define a function $\varphi(a,b) = a$, then,

$$T_1(\mathbf{x}) = \bar{x} = \varphi(\bar{x}, s^2) = \varphi(T_2(\mathbf{x})).$$

# Minimal sufficient statistics

***Sufficient condition***

Let $f(\mathbf{x} \mid \theta)$ be the pmf or pdf of a sample $\mathbf{X}$. Suppose there exists a function $T(\mathbf{x})$ such that, for every two sample points $\mathbf{x}$ and $\mathbf{y}$, the ratio $f(\mathbf{x} \mid \theta) \,/\, f(\mathbf{y} \mid \theta)$ is constant as a funciton of $\theta$ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Then $T(\mathbf{X})$ is a minimal sufficient statistic for $\theta$.

# Normal minimal sufficient statistic

Let $X_1,\ldots,X_n$ be iid random variables from a normal population $N(\mu,\sigma^2)$, both $\mu$, $\sigma^2$ are unknown. Let $\mathbf{x}$ and $\mathbf{y}$ denote two sample points, and let $(\overline{x}, s_x^2)$ and $(\overline{y}, s_y^2)$ be the sample means and variances corresponding to the sample points of $\mathbf{x}$ and $\mathbf{y}$, respectively.

Then,

$$f(\mathbf{x} \mid \mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{n}{2\sigma^2}(\overline{x}-\mu)^2 - \frac{n-1}{2\sigma^2}s_x^2\right],$$

$$f(\mathbf{y} \mid \mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{n}{2\sigma^2}(\overline{y}-\mu)^2 - \frac{n-1}{2\sigma^2}s_y^2\right],$$

$$\frac{f(\mathbf{x} \mid \mu,\sigma^2)}{f(\mathbf{y} \mid \mu,\sigma^2)} = \exp\left\{-\frac{1}{2\sigma^2}\left[-n(\overline{x}^2-\overline{y}^2) + 2n\mu(\overline{x}-\overline{y}) - (n-1)(s_x^2 - s_y^2)\right]\right\},$$

which will be constant as a function of $(\mu,\sigma^2)$ if and only if $\overline{x} = \overline{y}$ and $s_x^2 = s_y^2$.

Therefore, $(\bar{X},\ S^2)$ is a minimal sufficient statistic of $(\mu,\sigma^2)$.

# Normal minimal sufficient statistic

Since

$$
\begin{aligned}
(n-1)s^2 &= \sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}x_i^2 - n\bar{x}^2 \\
\frac{f(\mathbf{x}\mid\mu,\sigma^2)}{f(\mathbf{y}\mid\mu,\sigma^2)} &= \exp\left\{-\frac{1}{2\sigma^2}\left[-n(\bar{x}^2-\bar{y}^2)+2n\mu(\bar{x}-\bar{y})-(n-1)(s_x^2-s_y^2)\right]\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left[-n(\bar{x}^2-\bar{y}^2)+2n\mu(\bar{x}-\bar{y})-\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right)+\left(\sum_{i=1}^{n}y_i^2-n\bar{y}^2\right)\right]\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left[2n\mu(\bar{x}-\bar{y})-\left(\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}y_i^2\right)\right]\right\}
\end{aligned}
$$

which will be constant as a function of $(\mu,\sigma^2)$ if and only if $\bar{x}=\bar{y}$ and $\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}y_i^2$.

Therefore, $\left(\sum_{i=1}^{n}X_i,\sum_{i=1}^{n}X_i^2\right)$ is also a minimal sufficient statistic of $(\mu,\sigma^2)$.

**A minimal sufficient statistic is not unique. Any one-to-one function of a minimal sufficient statistic is also a minimal sufficient statistic.**

# Ancillary statistics

### *Ancillary statistics*

A statistic $S(\mathbf{X})$ whose distribution does not depend on the parameter $\theta$ is called an ancillary statistic.

Alone, an ancillary statistic contains **no** information about the parameter. When used in conjunction with other statistics, however, an ancillary statistic sometimes does contain valuable information for inferences about the parameter.

# Location family ancillary statistic

Let $X_1,\ldots,X_n$ be iid random variables from a location parameter family with cdf $F(x-\theta)$, $-\infty<\theta<\infty$. Let $Z_1=X_1-\theta,\ldots,Z_n=X_n-\theta$. We have that $Z_1,\ldots,Z_n$ are iid random variables from $F(x)$. Now, consider the range statistic $R=X_{(n)}-X_{(1)}$.

$$
\begin{aligned}
F_R(r\mid\theta) &= P(R\le r\mid\theta)\\
&= P\left(\max_{1\le i\le n} X_i-\min_{1\le i\le n} X_i\le r\mid\theta\right)\\
&= P\left(\max_{1\le i\le n}(Z_i+\theta)-\min_{1\le i\le n}(Z_i+\theta)\le r\mid\theta\right)\\
&= P\left(\max_{1\le i\le n} Z_i-\min_{1\le i\le n} Z_i+\theta-\theta\le r\mid\theta\right)\\
&= P\left(\max_{1\le i\le n} Z_i-\min_{1\le i\le n} Z_i\le r\mid\theta\right)\\
&= P\left(\max_{1\le i\le n} Z_i-\min_{1\le i\le n} Z_i\le r\right)
\end{aligned}
$$

**The range statistic is an ancillary statistic for the location parameter.**

# Scale family ancillary statistic

Let $X_1,\ldots,X_n$ be iid random variables from a location parameter family with cdf $F(x/\sigma)$, $\sigma > 0$. Let $Z_1 = X_1/\sigma,\ldots,Z_n = X_n/\sigma$. We have that $Z_1,\ldots,Z_n$ are iid random variables from $F(x)$. Now, consider the statistic $T(\mathbf{X}) = (X_1/X_n,\ldots,X_{n-1}/X_n)$. Let $Y_i = X_i/X_n$. Then,

$$\begin{aligned}
F(y_1,\ldots,y_{n-1}\mid\sigma) &= P(Y_1 \le y_1,\ldots,Y_{n-1} \le y_{n-1}\mid\sigma) \\
&= P\left(X_1/X_n \le y_1,\ldots,X_{n-1}/X_n \le y_{n-1}\mid\sigma\right) \\
&= P\left((\sigma Z_1/\sigma Z_n) \le y_1,\ldots,(\sigma Z_{n-1}/\sigma Z_n) \le y_{n-1}\mid\sigma\right) \\
&= P\left(Z_1/Z_n \le y_1,\ldots,Z_{n-1}/Z_n \le y_{n-1}\mid\sigma\right) \\
&= P\left(Z_1/Z_n \le y_1,\ldots,Z_{n-1}/Z_n \le y_{n-1}\right)
\end{aligned}$$

**Any statistic that depends on the sample only through the $n$-1 values $X_1/X_n$, …, $X_{n-1}/X_n$ is an ancillary statistic for the scale parameter.**

# Complete statistics

## *Complete statistics*

Let $f(t \mid \theta)$ be a family of pdfs or pmfs for a statistic $T(\mathbf{X})$. The family of probability distribution is called complete if $\mathrm{E}_\theta g(T) = 0$ for all $\theta$ implies $P_\theta(g(T) = 0) = 1$ for all $\theta$. Equivalently, $T(\mathbf{X})$ is called a complete statistic.

# Binomial complete sufficient statistic

Suppose that $T$ has a binomial$(n,p)$ distribution, $0<p<1$. Let $g(\cdot)$ be a function such that $\mathrm{E}_p g(T)=0$. Then,

$$\mathrm{E}_p g(T)=\sum_{t=0}^{n} g(t)\binom{n}{t}p^t(1-p)^{n-t}=(1-p)^n\sum_{t=0}^{n} g(t)\binom{n}{t}\left(\frac{p}{1-p}\right)^t$$

In order to ensure $\mathrm{E}_p g(T)=0$ for all $0<p<1$, $g(t)$ must be 0 for all $t$.

In other words,

$$P_p(g(T)=0)=1.$$

Therefore, $T$ is a complete statistic.

**The probability that $g(T)$=0 must be 1.**

# Exponential family

**_Complete statistic in the exponential family_**

Let $X_1, \ldots, X_n$ be iid random variables from a pdf or pmf that belongs to an exponential family given by

$$f(x \mid \boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left[\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right],$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d), d \leq k$. Then,

$$T(X) = \left(\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is a complete statistic as long as the parameter space $\Theta$ contains an open set in $\Re^k$.

# Basu's theorem

***Basu's theorem***

If $T(\mathbf{X})$ is a complelte and minimal sufficient statistic, then $T(\mathbf{X})$ is independent of every ancillary statistic.

If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.

# Normal complete statistic

Let $X_1,\ldots,X_n$ be iid random variables with common pdf $N(\mu,\sigma^2)$, where $\mu$ is unknown but $\sigma^2$ is known. Then,

$\bar{X}$ is a sufficient statistic for $\mu$.

$\bar{X}$ is a minimal sufficient statistic for $\mu$.

$\bar{X}$ is a complete statistic.

$S^2$ is an ancillary statistic for $\mu$.

By Basu's theorem,

The complete and minimal sufficient statistic $\bar{X}$ is independent of the ancillary statistic $S^2$.

# The likelihood principle

Let $X_1, \ldots, X_n$ be iid random variables from a Bernoulli $(\theta)$ population.

Then the joint pdf of $X_1, \ldots, X_n$ is

$$f(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i} = \theta^{n_1}(1-\theta)^{n-n_1},$$

where, $n_1 = \sum_{i=1}^{n} x_i$.

Now, we have two guesses, $\theta_1$ and $\theta_2$, about the true parameter $\theta$.
Which one is more likely to be true?

Since $f(\mathbf{x} \mid \theta) = P(\mathbf{X} = \mathbf{x} \mid \theta)$, we may like to compare the two probabilities

$$f(\mathbf{x} \mid \theta_1) \text{ vs. } f(\mathbf{x} \mid \theta_2).$$

If $f(\mathbf{x} \mid \theta_1) > f(\mathbf{x} \mid \theta_2)$, $\theta_1$ is more likely to be true.

If $f(\mathbf{x} \mid \theta_1) = f(\mathbf{x} \mid \theta_2)$, $\theta_1$ and $p_2$ are equally likely to be true.

If $f(\mathbf{x} \mid \theta_1) < f(\mathbf{x} \mid \theta_2)$, $\theta_2$ is more likely to be true.

# The likelihood function

***Likelihood function***

Let $f(\mathbf{x} \mid \theta)$ denote the joint pmf or pdf of the sample $\mathbf{X} = (X_1, \ldots X_n)$. Then, given that $\mathbf{X} = \mathbf{x}$ is observed, the function of $\theta$ defined by

$$L(\theta \mid \mathbf{x}) = f(\mathbf{x} \mid \theta)$$

is called the **likelihood function**.

The likelihood function measures the plausibility that the sample is observed under a certain parameter. Larger likelihood means the sample that we observed is more likely to have occurred due to the given parameter.

# Bernoulli likelihood function

Let $X_1, \ldots, X_n$ be iid random variables from a Bernoulli $(\theta)$ population.
Then the joint pdf of $X_1, \ldots, X_n$ is

$$f(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n-\sum_{i=1}^{n} x_i} = \theta^{n_1} (1-\theta)^{n-n_1},$$

where $n_1 = \sum_{i=1}^{n} x_i$.

Therefore, the likelihood function for $p$ is given by

$$L(\theta|\mathbf{x}) = \theta^{n_1} (1-\theta)^{n-n_1}.$$

In $f(\mathrm{x}|\theta)$, $\theta$ is fixed, and x is varying over all possible sample points.
In $L(\theta|\mathrm{x})$, however, x is fixed, and $\theta$ is varying over all possible parameter values.

# Normal likelihood function

Let $X_1,\dots,X_n$ be iid random variables from a normal population $N(\mu,\sigma^2)$, where $\sigma^2$ is already know and the only parameter is $\mu$. Then the joint pdf of $X_1,\dots,X_n$ is

$$f(\mathbf{x} \mid \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right],$$

Therefore, the likelihood function for $\mu$ is given by

$$L(\mu|\mathbf{x}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right].$$

In $f(\mathrm{x}|\mu)$, $\mu$ is fixed, and x is varying over all possible sample points.
In $L(\mu|\mathrm{x})$, however, x is fixed, and $\mu$ is varying over all possible parameter values.

# Normal likelihood function

Let $X_1,\ldots,X_n$ be iid random variables from a normal population $N(\mu,\sigma^2)$.

Then the joint pdf of $X_1,\ldots,X_n$ is

$$f(\mathbf{x} \mid \mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right],$$

Therefore, the likelihood function for $(\mu,\sigma^2)$ is given by

$$L(\mu,\sigma^2|\mathbf{x}) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right].$$

In $f(\mathrm{x}|\mu,\ \sigma^2)$, $(\mu,\ \sigma^2)$ is fixed, and x is varying over all possible sample points.
In $L(\mu,\ \sigma^2|\mathrm{x})$, however, x is fixed, and $(\mu,\ \sigma^2)$ is varying over all possible parameter values.

# Normal likelihood function

Let $X_1,\ldots,X_n$ be iid random variables from a normal population $N(\mu,\sigma^2)$, where $\sigma^2$ is already known and the only parameter is $\mu$. From previoius results, we know that $\bar{X}$ is a sufficient statistic of $\mu$, and more importantly, $\bar{X} \sim N(\mu,\sigma^2/n)$.

Then the pdf of $\bar{X}$ is

$$f(\overline{x} \mid \mu) = \frac{1}{\sqrt{2\pi}\sigma/\sqrt{n}} \exp\left[-\frac{(\overline{x}-\mu)^2}{2\sigma^2/n}\right],$$

and the likelihood function for $\mu$ is given by

$$L(\mu \mid \overline{x}) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left[-\frac{n(\overline{x}-\mu)^2}{2\sigma^2}\right].$$

# Calculations of likelihoods

**Suppose that 1000 Bernoulli trials have been done, $n$=1000, $n_1$=500. Then the likelihood for $p$=0.5 is**

$$0.5^{500}(1-0.5)^{1000-500} = 0.5^{1000} \approx 9.33\times 10^{-302}$$

**Suppose that 800 observations have been obtained from a standard normal population, and their squares add up to 800. Then the likelihood for ($\mu$, $\sigma^2$)=(0, 1) is**

$$(2\pi)^{-400}e^{-400} \approx 5.35\times 10^{-320}\times 1.92\times 10^{-174} \approx 1.03\times 10^{-493}$$

# Log likelihoods

Let $X_1, \ldots, X_n$ be iid random variables from a Bernoulli $(\theta)$ population.
Then the likelihood function for $\theta$ is

$$L(\theta|\mathbf{x}) = \theta^{n_1}(1-\theta)^{n-n_1}.$$

Therefore, the log likelihood is

$$l(\theta|\mathbf{x}) = \log L(\theta|\mathbf{x}) = n_1 \log\theta + (n - n_1)\log(1-\theta).$$

Let $X_1, \ldots, X_n$ be iid random variables from a normal $(\mu, \sigma^2)$ population.
Then the likelihood function for $(\mu, \sigma^2)$ is

$$L(\mu,\sigma^2|\mathbf{x}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right].$$

Therefore, the log likelihood is

$$l(\mu,\sigma^2|\mathbf{x}) = \log L(\mu,\sigma^2|\mathbf{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}(\sigma^2)^{-1}\sum_{i=1}^{n}(x_i-\mu)^2.$$

# Likelihood ratio

Let $X_1,\ldots,X_n$ be iid random variables from a multinomial trial population with cell probability $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_m)$ . Then the joint pdf of $X_1,\ldots,X_n$ is

$$L(\boldsymbol{\theta}|\mathbf{x}) = \prod_{j=1}^{m} \theta_j^{n_j} \text{ and } l(\boldsymbol{\theta}|\mathbf{x}) = L(\boldsymbol{\theta}|\mathbf{x}) = \sum_{j=1}^{m} n_j \log \theta_j,$$

where $n_j = \sum_{i=1}^{n} I(x_i = j), j = 1,\ldots m.$

Suppose that we have two guesses for $\boldsymbol{\theta}$, say, $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$. Then,

$$\frac{L(\boldsymbol{\theta}^{(1)}|\mathbf{x})}{L(\boldsymbol{\theta}^{(2)}|\mathbf{x})} = \frac{\prod_{j=1}^{m} \left(\theta_j^{(1)}\right)^{n_j}}{\prod_{j=1}^{m} \left(\theta_j^{(2)}\right)^{n_j}} = \prod_{j=1}^{m} \left(\frac{\theta_j^{(1)}}{\theta_j^{(2)}}\right)^{n_j}.$$

Obvioiusly,

$$\log \frac{L(\boldsymbol{\theta}^{(1)}|\mathbf{x})}{L(\boldsymbol{\theta}^{(2)}|\mathbf{x})} = l(\boldsymbol{\theta}^{(1)}|\mathbf{x}) - l(\boldsymbol{\theta}^{(2)}|\mathbf{x}) = \sum_{j=1}^{m} n_j \left(\log \theta_j^{(1)} - \log \theta_j^{(2)}\right).$$

# Likelihood ratio for comparing parameters

Intuitively, the likelihood ratio provides a means of measuring the goodness of $\theta^{(1)}$ and $\theta^{(2)}$.

If $L(\theta^{(1)}|\mathbf{x})/L(\theta^{(2)}|\mathbf{x}) > 1$, $\theta^{(1)}$ is more likely to be the true.

If $L(\theta^{(1)}|\mathbf{x})/L(\theta^{(2)}|\mathbf{x}) = 1$, $\theta^{(1)}$ and $\theta^{(2)}$ are equally likely to be true.

If $L(\theta^{(1)}|\mathbf{x})/L(\theta^{(2)}|\mathbf{x}) < 1$, $\theta^{(2)}$ is more likely to be the true.

But how about we have another sample point $\mathbf{y}$ instead of $\mathbf{x}$, in what condition we would have the same inference results?

# The likelihood principle

**LIKELIHOOD PRINCIPLE**

If $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that $L(\theta \mid \mathbf{x})$ is proportional to $L(\theta \mid \mathbf{y})$, that is, there exists a constant $C(\mathbf{x}, \mathbf{y})$ such that

$$L(\theta \mid \mathbf{x}) = C(\mathbf{x}, \mathbf{y}) L(\theta \mid \mathbf{y}) \quad \text{for all } \theta,$$

then the conclusions drawn from $\mathbf{x}$ and $\mathbf{y}$ should be identical.

$$\frac{L(\theta^{(1)} \mid \mathbf{x})}{L(\theta^{(2)} \mid \mathbf{x})} = \frac{C(\mathbf{x}, \mathbf{y}) L(\theta^{(1)} \mid \mathbf{y})}{C(\mathbf{x}, \mathbf{y}) L(\theta^{(2)} \mid \mathbf{y})} = \frac{L(\theta^{(1)} \mid \mathbf{y})}{L(\theta^{(2)} \mid \mathbf{y})}$$

# Descriptive Statistics

统计学方法及其应用 | 统计学基础 | 随机变量的函数

*"A random variable is a quantity whose values are random and to which a probability distribution is assigned."*

# Summary statistics

**数据的概括**

数据的概括就是根据数据简约的原理，设计出描述统计量来描述试验数据。

我们处理的是样本的观测值！

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

自学 *Introductory statistics with R*, page 57-80

# Quantiles

- 四分位数 (Quartiles)
  - 1st quartile, Median, 3rd quartile
  - Interquartile range (IQR)
- 十分位数 (Centiles)
- 百分位数 (Percentiles)

```
> quantile(x)
> quantile(x, seq(0, 1, 0.1))
> quantile(x, seq(0, 1, 0.01))
> quantile(x, type=2)
```

| Order | Value |
|---|---|
| (1) | Min |
| (2) | |
| (3) | |
| (4) | |
| (5) | 1st Qu. |
| (6) | |
| (7) | |
| (8) | |
| (9) | |
| (10) | Median |
| (11) | |
| (12) | |
| (13) | |
| (14) | |
| (15) | 3rd Qu. |
| (16) | |
| (17) | |
| (18) | |
| (19) | Max |

IQR

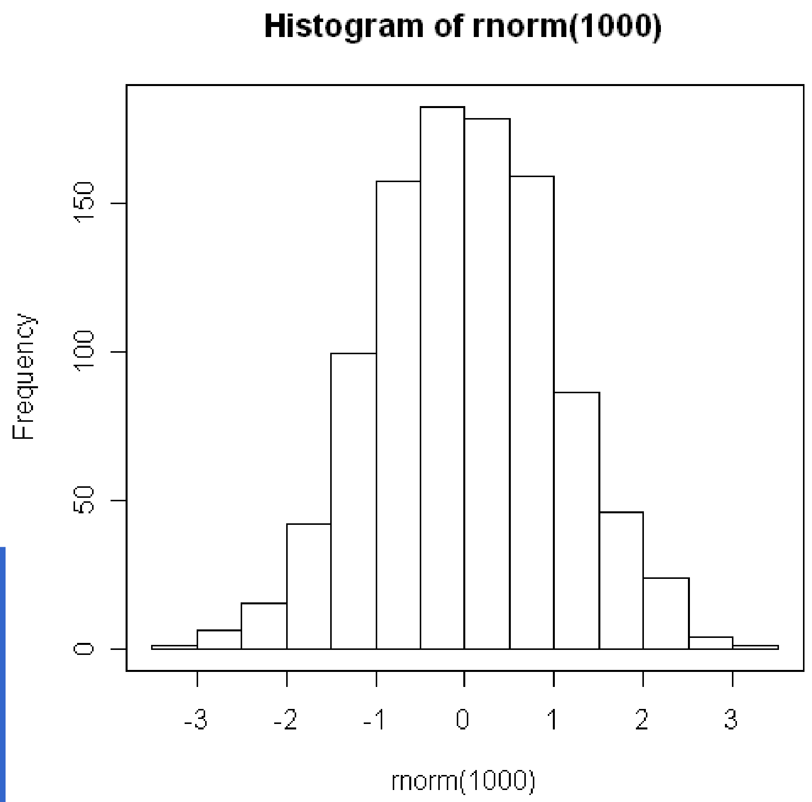
# Summary statistics

```
> fivenum(x)
> summary(x)
```

| Index | Statistic |
| --- | --- |
| 1 | Min |
| 2 | 1st Qu. |
| 3 | Median |
| 4 | Mean |
| 5 | 3rd Qu. |
| 6 | Max |
| 7 | NAs |

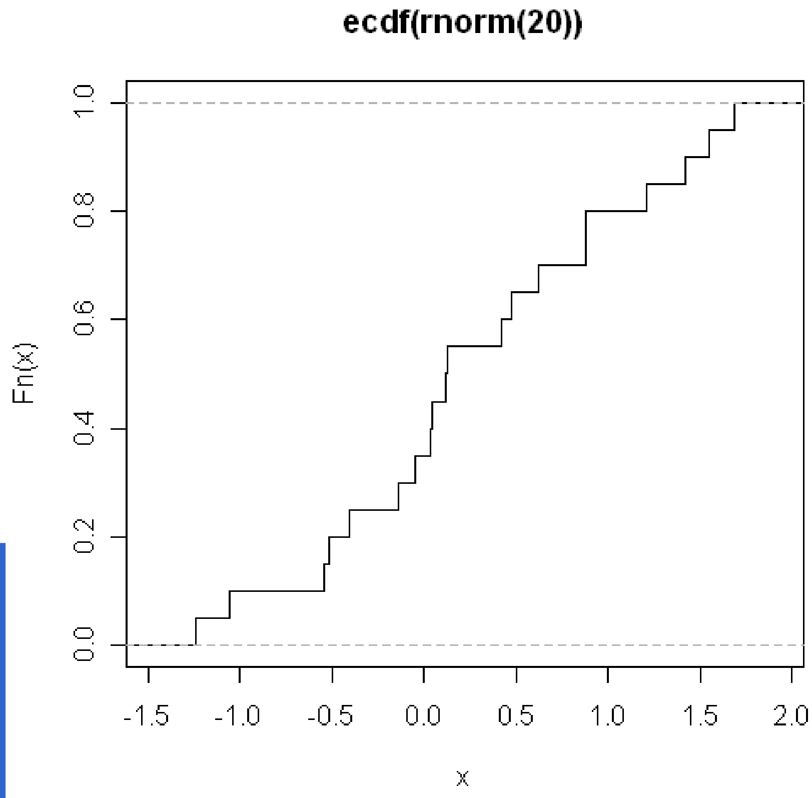
# Histograms

```
> hist(x)
> hist(x, freq=F)
> hist(x, freq=F, col="red")
> H <- hist(x)
```

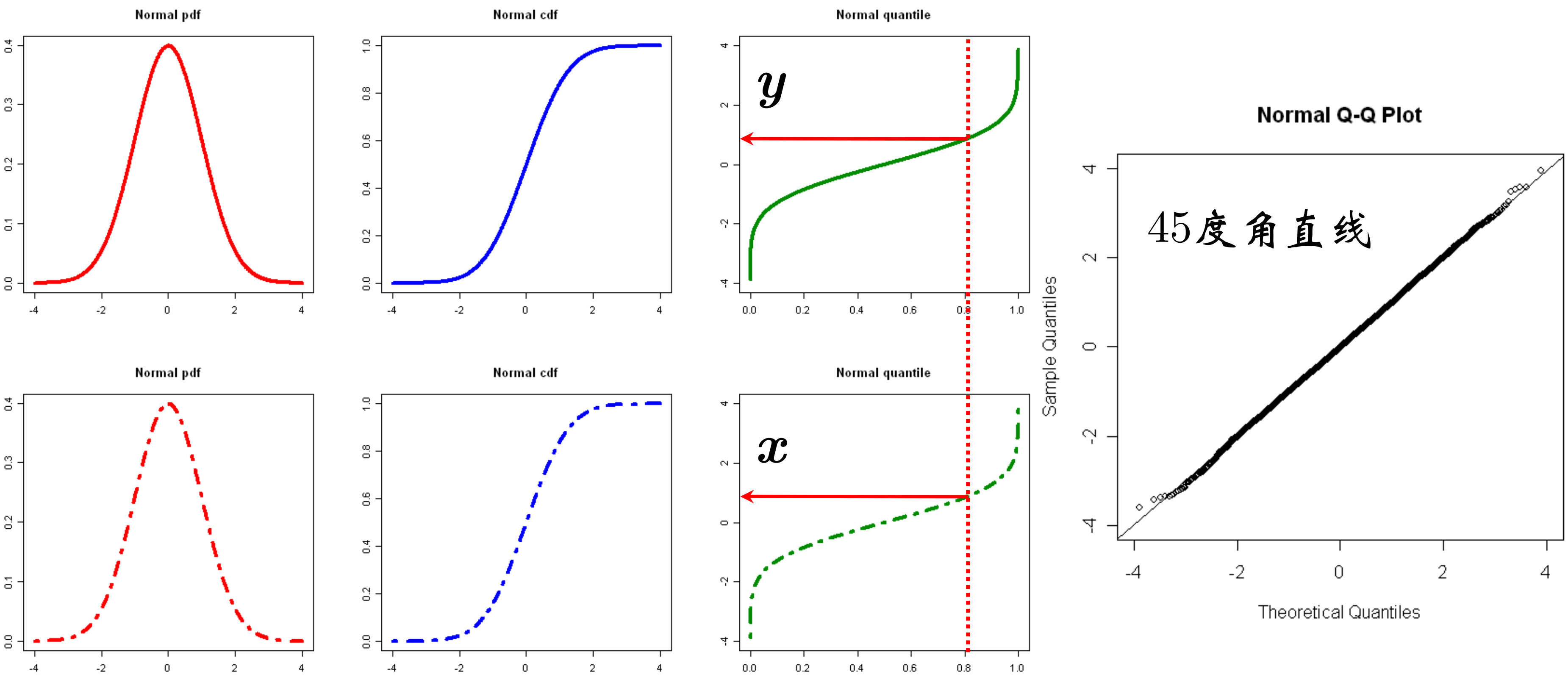


# Empirical cdf (ecdf)

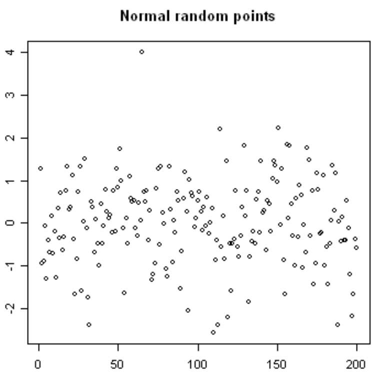
```
> F <- ecdf(x)
> plot(ecdf(x))
```

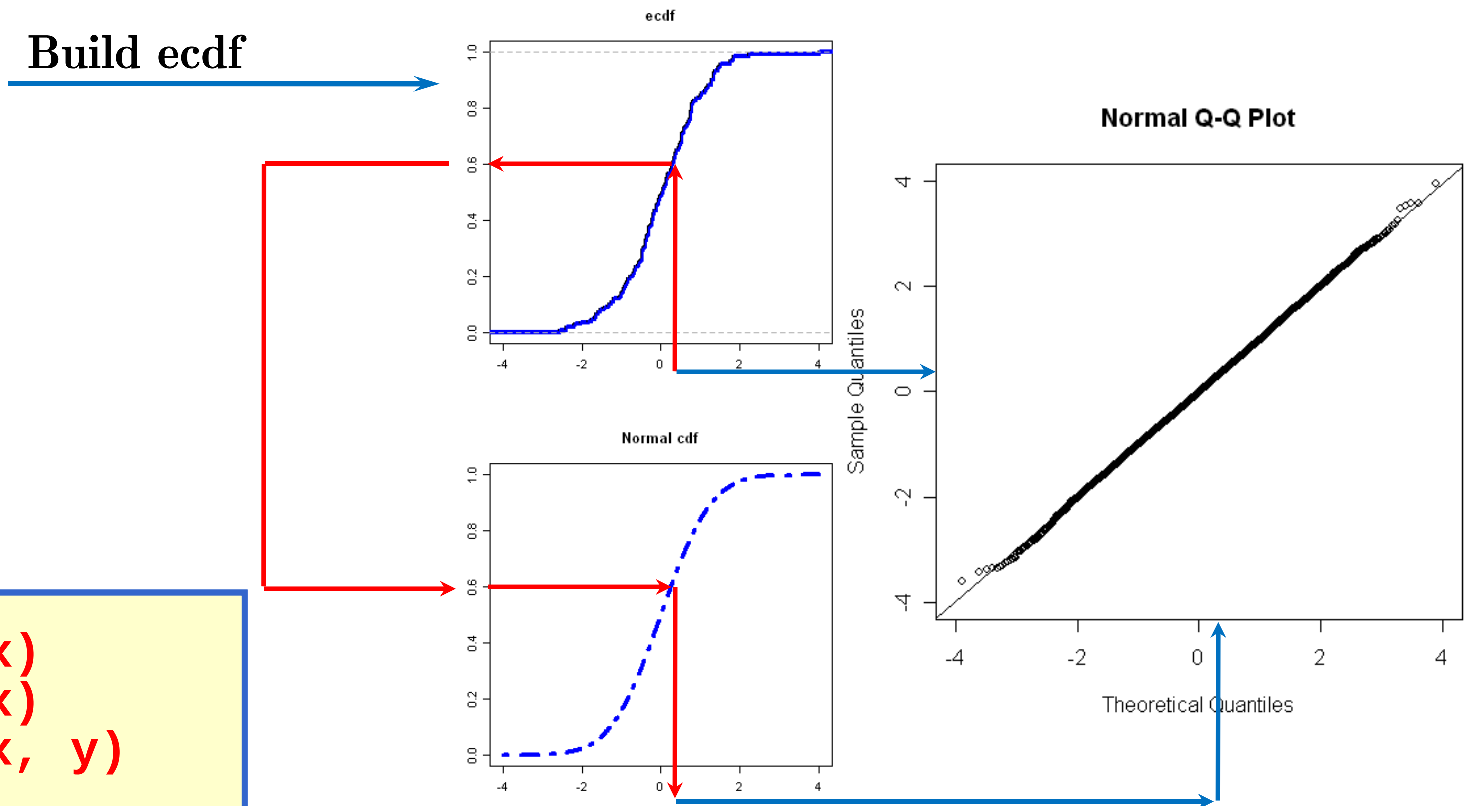


# Q-Q plots


Normal pdf
Normal cdf
Normal quantile
y
Normal pdf
Normal cdf
Normal quantile
x
Normal Q-Q Plot
45度角直线
Sample Quantiles
Theoretical Quantiles

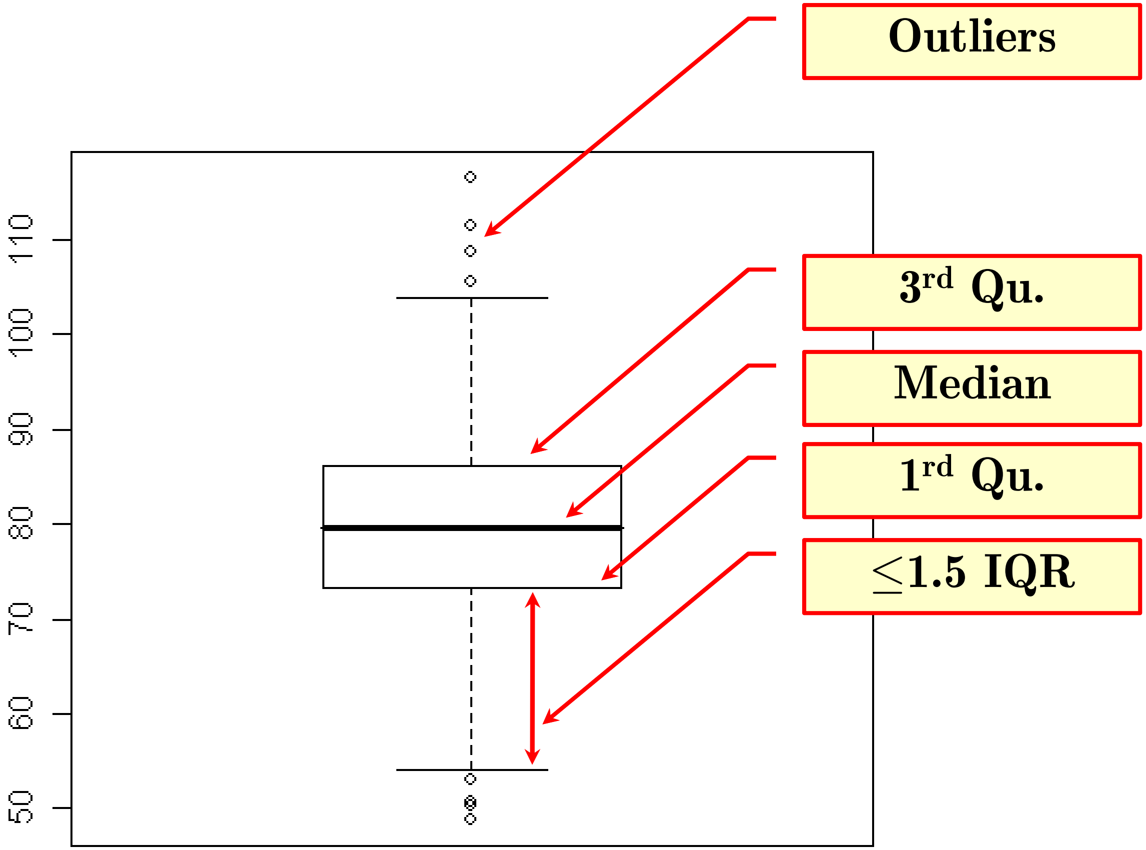
# How to build a Q-Q plot




```
> qqnorm(x)
> qqline(x)
> qqplot(x, y)
```

# Box plots




```
> boxplot(x)
> boxplot.stats(x)
```

# Inferential Statistics

统计学方法及其应用 | 统计学基础 | 随机变量的函数

*"A random variable is a quantity whose values are random and to which a probability distribution is assigned."*

# Inferential statistics

- 点估计 (Point estimation)
- 假设检验 (Hypothesis testing)
- 区间估计 (Interval estimation)
- 方差分析 (Analysis of variance)
- 回归分析 (Regression models)

# Point Estimation



*"A random variable is a quantity whose values are random and to which a probability distribution is assigned."*

# Introduction

- For a parametric model

$$f(x \mid \theta)$$

- The mathematical structure is already know
- The knowledge of the parameter yield the knowledge of the entire population
- We are interested in obtaining a good estimation of $\theta$
  Sometimes an estimation of a function of $\theta$

# Point estimator

## *Point estimator*

A **point estimator** is any function $W(X_1,\ldots,X_n)$ of a sample; that is, any statistic is a point estimator.

**Estimator:** a function of the sample, a random variable.

$$W(X_1,\ldots,X_n)$$

**Estimate:** the realized value of an estimator, a number.

$$W(x_1,\ldots,x_n)$$

# Say "NO" to trivial things

$$W(X_1, \ldots, X_n) = 3.14159265358979323846$$

# Method of Moments



*"A random variable is a quantity whose values are random and to which a probability distribution is assigned."*

# Method of moments

Let $X_1, \ldots, X_n$ be a sample from a population with $k$ parameters $f(x \mid \theta_1, \ldots, \theta_k)$. Define

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i^1, \qquad \mu'_1 = \mathrm{E}X^1;$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \qquad \mu'_2 = \mathrm{E}X^2;$$

$$\ldots$$

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \qquad \mu'_k = \mathrm{E}X^k.$$

Solve the system of equations for $(\theta_1, \ldots, \theta_k)$, in terms of $(m_1, \ldots, m_k)$

$$m_1 = \mu'_1(\theta_1, \ldots, \theta_k);$$

$$m_2 = \mu'_2(\theta_1, \ldots, \theta_k);$$

$$\ldots$$

$$m_k = \mu'_k(\theta_1, \ldots, \theta_k).$$

# Bernoulli method of moments

Let $X_1, \ldots, X_n$ be a sample from a Bernoulli $(\theta)$ population. Then,

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}, \qquad \mu'_1 = \theta;$$

Solve

$$\bar{X} = \theta.$$

We have

$$\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

# Normal method of moments

Let $X_1, \ldots, X_n$ be a sample from a normal population $\mathrm{N}(\mu, \sigma^2)$.
Then,

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}, \qquad \mu'_1 = \mu;$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \qquad \mu'_2 = \mu^2 + \sigma^2.$$

Solve

$$\begin{cases} \overline{X} = \mu \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2 = \mu^2 + \sigma^2 \end{cases}.$$

We have
$$\begin{cases} \hat{\mu} = \bar{X} \\ \hat{\sigma}^2 = \dfrac{1}{n}\left(\displaystyle\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})^2 = \dfrac{n-1}{n} S^2 \end{cases}.$$

# Satterthwaite approximation

Let $Y_i, i = 1, \ldots, k$ be independent random variables, with

$$Y_i \sim \chi^2_{r_i}$$

Then

$$\sum_{i=1}^{k} Y_i \sim \chi^2_{\sum_{i=1}^{k} r_i}$$

But, how about

$$\sum_{i=1}^{k} a_i Y_i$$

where $a_i$s are known constants.

# Background — I

$X_1,\ldots,X_m$, a sample from $N(\mu_X,\sigma_X^2)$, $Y_1,\ldots,Y_n$, a sample from $N(\mu_Y,\sigma_Y^2)$

Then,

$$\bar{X} \sim N(\mu_X,\sigma_X^2 / m),\ \ \bar{Y} \sim N(\mu_Y,\sigma_Y^2 / n)$$

Thus,

$$\bar{X} - \bar{Y} \sim N(\mu_X - \mu_Y,\sigma_X^2 / m + \sigma_Y^2 / n)$$

Or equivalently,

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2 / m + \sigma_Y^2 / n}} \sim N(0,1)$$

Now,

both $\sigma_X^2$ and $\sigma_Y^2$ are unknown and assume $\sigma_X^2 \neq \sigma_Y^2$. What is the distribution of

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{S_X^2 / m + S_Y^2 / n}}$$

# Background — II

The ideal form is

$$\frac{S_X^2 / m + S_Y^2 / n}{\sigma_X^2 / m + \sigma_Y^2 / n} \sim \frac{\chi_v^2}{v}$$

in order to apply our privious results (student's $t$ distribution).

Because

$$\frac{S_X^2 / m}{\sigma_X^2 / m + \sigma_Y^2 / n} = \underbrace{\left( \frac{1}{m(m-1)} \frac{\sigma_X^2}{\sigma_X^2 / m + \sigma_Y^2 / n} \right)}_{a_X} \underbrace{\frac{(m-1)S_X^2}{\sigma_X^2}}_{\chi_{m-1}^2}$$

$$\frac{S_Y^2 / n}{\sigma_X^2 / m + \sigma_Y^2 / n} = \underbrace{\left( \frac{1}{n(n-1)} \frac{\sigma_Y^2}{\sigma_X^2 / m + \sigma_Y^2 / n} \right)}_{a_Y} \underbrace{\frac{(n-1)S_Y^2}{\sigma_Y^2}}_{\chi_{n-1}^2}$$

We meet the Satterthwaite approximation problem.

# Naïve way

If the approximation $\sum_{i=1}^{k} a_i Y_i \sim \chi_v^2 / v$ holds, we have

$$\mathrm{E}\left(\sum_{i=1}^{k} a_i Y_i\right) = \mathrm{E}\left(\chi_v^2 / v\right) = \mathrm{E}\left(\chi_v^2\right) / v = 1, \text{ and}$$

$$\mathrm{E}\left(\sum_{i=1}^{k} a_i Y_i\right)^2 = \mathrm{E}\left(\chi_v^2 / v\right)^2 = \frac{1}{v^2}\left[\mathrm{Var}\chi_v^2 + \left(\mathrm{E}\chi_v\right)^2\right] = \frac{2}{v} + 1$$

Now,

$$\sum_{i=1}^{k} a_i Y_i = 1 \qquad \Rightarrow \text{No information}$$

$$\left(\sum_{i=1}^{k} a_i Y_i\right)^2 = \frac{2}{v} + 1 \qquad \Rightarrow v = \frac{2}{\left(\sum_{i=1}^{k} a_i Y_i\right)^2 - 1}$$

Therefore

$$\hat{v} = \frac{2}{\left(\sum_{i=1}^{k} a_i Y_i\right)^2 - 1}$$

# Solution

If the approximation $\sum_{i=1}^{k} a_i Y_i \sim \chi_v^2 / v$ holds, we have

$$\mathrm{E}\left(\sum_{i=1}^{k} a_i Y_i\right) = \mathrm{E}\left(\chi_v^2 / v\right) = \mathrm{E}\left(\chi_v^2\right) / v = 1, \text{ and}$$

Now,

$$\mathrm{E}\left(\sum_{i=1}^{k} a_i Y_i\right)^2 = \mathrm{E}\left(\chi_v^2 / v\right)^2 = \frac{1}{v^2}\left[\mathrm{Var}\chi_v^2 + \left(\mathrm{E}\chi_v\right)^2\right] = \frac{2}{v} + 1$$

$$\mathrm{E}\left(\sum_{i=1}^{k} a_i Y_i\right)^2 = \mathrm{Var}\left(\sum_{i=1}^{k} a_i Y_i\right) + \left(\mathrm{E}\sum_{i=1}^{k} a_i Y_i\right)^2 = \underbrace{\left(\mathrm{E}\sum_{i=1}^{k} a_i Y_i\right)^2}_{1}\left[\frac{\mathrm{Var}\left(\sum_{i=1}^{k} a_i Y_i\right)}{\left(\mathrm{E}\sum_{i=1}^{k} a_i Y_i\right)^2} + 1\right]$$

We have

$$v = \frac{2\left(\mathrm{E}\sum_{i=1}^{k} a_i Y_i\right)^2}{\mathrm{Var}\left(\sum_{i=1}^{k} a_i Y_i\right)} = \frac{2\left(\sum_{i=1}^{k} a_i \mathrm{E}\, Y_i\right)^2}{\sum_{i=1}^{k} a_i^2 \mathrm{Var}\, Y_i} = \frac{2\left(\sum_{i=1}^{k} a_i \mathrm{E}\, Y_i\right)^2}{2\sum_{i=1}^{k} (a_i^2 / r_i)\left(\mathrm{E}\, Y_i\right)^2}$$

Therefore

$$\hat{v} = \frac{\left(\sum_{i=1}^{k} a_i Y_i\right)^2}{\sum_{i=1}^{k} (a_i^2 / r_i)\, Y_i^2}$$

$\mathrm{E}\left(\chi_p^2\right) = p$

$\mathrm{Var}\left(\chi_p^2\right) = 2p = 2p^2 / p = 2\left[\mathrm{E}\left(\chi_p^2\right)\right]^2 / p$

$\mathrm{E}\left(Y_i\right) = Y_i$, method of moments, $n = 1$

# Return back to our problem

When approximating

$$\frac{S_X^2 / m + S_Y^2 / n}{\sigma_X^2 / m + \sigma_Y^2 / n} \sim \frac{\chi_\upsilon^2}{\upsilon}$$

We need to choose

$$\hat{\upsilon} = \frac{\left(\dfrac{S_X^2}{m} + \dfrac{S_Y^2}{n}\right)^2}{\dfrac{S_X^4}{m^2(m-1)} + \dfrac{S_Y^4}{n^2(n-1)}}$$

With this approximation,

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{S_X^2 / m + S_Y^2 / n}}$$

will have a student's t distribution with $\hat{\upsilon}$ degrees of freedom.

# Maximum Likelihood Estimation

统计学方法及其应用 统计学基础 随机变量的函数

*"A random variable is a quantity whose values are random and to which a probability distribution is assigned."*

# Likelihood ratio

Intuitively, the likelihood ratio provides a means of measuring the goodness of $\theta^{(1)}$ and $\theta^{(2)}$.

If $L(\theta^{(1)}|\mathbf{x})/L(\theta^{(2)}|\mathbf{x}) > 1$, $\theta^{(1)}$ is more likely to be the true.

If $L(\theta^{(1)}|\mathbf{x})/L(\theta^{(2)}|\mathbf{x}) = 1$, $\theta^{(1)}$ and $\theta^{(2)}$ are equally likely to be true.

If $L(\theta^{(1)}|\mathbf{x})/L(\theta^{(2)}|\mathbf{x}) < 1$, $\theta^{(2)}$ is more likely to be the true.

But how about we have another sample point $\mathbf{y}$ instead of $\mathbf{x}$, in what condition we would have the same inference results?

$$\frac{L(\theta^{(1)}|\mathbf{x})}{L(\theta^{(2)}|\mathbf{x})} = \frac{L(\theta^{(1)}|\mathbf{y})}{L(\theta^{(2)}|\mathbf{y})} = \frac{const}{const}\frac{L(\theta^{(1)}|\mathbf{y})}{L(\theta^{(2)}|\mathbf{y})}$$

# Maximum likelihood estimate

Because a larger likelihood implies a bigger plausibility that a parameter is the true one. It is reasonable to choose the parameter $\theta^\star$ that can maximize the likelihood function $L(\theta \mid \mathbf{x})$ as our best guess of $\theta$.

In other words,

$$\theta^\star = \arg\max_{\theta \in \Theta} L(\theta \mid \mathbf{x}).$$

Equivalently,

$$\theta^\star = \arg\max_{\theta \in \Theta} \log L(\theta \mid \mathbf{x}).$$

Obviously,

$$L(\theta^\star \mid \mathbf{x}) \geq L(\theta \mid \mathbf{x}), \text{ for any } \theta \in \Theta.$$

$\theta^\star$ is called the *maximum likelihood estimate* (MLE) of $\theta$.

# Maximum likelihood estimators

***Maximum likelihood estimator***

For each sample point $\mathbf{x}$, let $\hat{\theta}(\mathbf{x})$ be a parameter value at which $L(\theta \mid \mathbf{x})$ attains its maximum as a function of $\theta$, with $\mathbf{x}$ held fixed. A **maximum likelihood estimator** (MLE) of the parameter $\theta$ based on a sample $\mathbf{X}$ is $\hat{\theta}(\mathbf{X})$.

We need to find a **global** maximum!
Need to check boundary conditions!
Sometimes yielding optimization problems with constraints.

**Refer to optimization books!**

# Normal MLE, mean

Let $X_1, \ldots, X_n$ be a sample from a normal population $\mathrm{N}(\mu, \sigma^2)$, where $\mu$ is unknown but $\sigma^2$ is known. Then, the likelihood function for $\mu$ is

$$L(\mu|\mathbf{x}) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right].$$

So we like to solve the optimization problem of

$$\max\ L(\mu|\mathbf{x}),$$

where $-\infty < \mu < \infty$. Let

$$\frac{d}{d\mu}L(\mu \mid \mathbf{x}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]\left[\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)\right] = 0,$$

we have $\displaystyle\sum_{i=1}^{n}(x_i - \mu) = 0.$

Therefore, $\displaystyle\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x},$

and $\displaystyle L(\hat{\mu} \mid \mathbf{x}) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]$

# Normal MLE, mean

Let $X_1, \ldots, X_n$ be a sample from a normal population $N(\mu, \sigma^2)$, where $\mu$ is unknown but $\sigma^2$ is known. Then, the log likelihood function for $\mu$ is

$$l(\mu|\mathbf{x}) = \log L(\mu|\mathbf{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

So we like to solve the optimization problem of

$$\max\ l(\mu|\mathbf{x}),$$

where $-\infty < \mu < \infty$. Let

$$\frac{d}{d\mu}l(\mu|\mathbf{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}x_i - n\mu\right) = 0,$$

we have

$$\sum_{i=1}^{n}x_i - n\mu = 0.$$

Therefore,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i = \overline{x}.$$

# Further check

Obvioiusly, $\hat{\mu} = \overline{x}$ is the only zero of the first order derivative.

Furthermore, for

$$l(p|\mathbf{x}) = \log L(p|\mathbf{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

$$\frac{d}{d\mu}l(\mu|\mathbf{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}x_i - n\mu\right),$$

$$\frac{d^2}{d\mu^2}l(\mu|\mathbf{x}) = -\frac{n}{\sigma^2} < 0,$$

Therefore,

$\hat{\mu} = \overline{x}$ is the only extreme point, and it is a maximum.

Since $\lim\limits_{\mu\to\infty} l(\mu|\mathbf{x}) = -\infty$ and $\lim\limits_{\mu\to-\infty} l(\mu|\mathbf{x}) = -\infty$,

$\hat{\mu} = \overline{x}$ is the only global maximum.

Hence,

$\overline{X}$ is the maximum likelihood estimator of $\mu$.

# Restricted Normal mean MLE

Let $X_1,\ldots,X_n$ be a sample from a normal population $N(\mu,\sigma^2)$, where $\sigma^2$ is known, $\mu$ is unknown but should satisfy $\mu=\mu_0$.

Needless to say, the single point $\mu_0$ itself is the MLE.

Therefore

$$\hat{\mu}=\mu_0,$$

and

$$L(\hat{\mu}\mid\mathbf{x})=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_0)^2\right].$$

# Restricted Normal mean MLE

Let $X_1,\dots,X_n$ be a sample from a normal population $\mathrm{N}(\mu,\sigma^2)$, where $\sigma^2$ is known, $\mu$ is unknown but should satisfy $\mu \ge \mu_0$.

The log likelihood function for $\mu$ is

$$l(\mu\,|\,\mathbf{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 - \frac{n}{2\sigma^2}(\mu-\bar{x})^2.$$

So we like to solve the optimization problem of

$$\begin{aligned} \max \quad & l(\mu\,|\,\mathbf{x}) \\ \text{s.t.} \quad & \mu \ge \mu_0 \end{aligned}$$

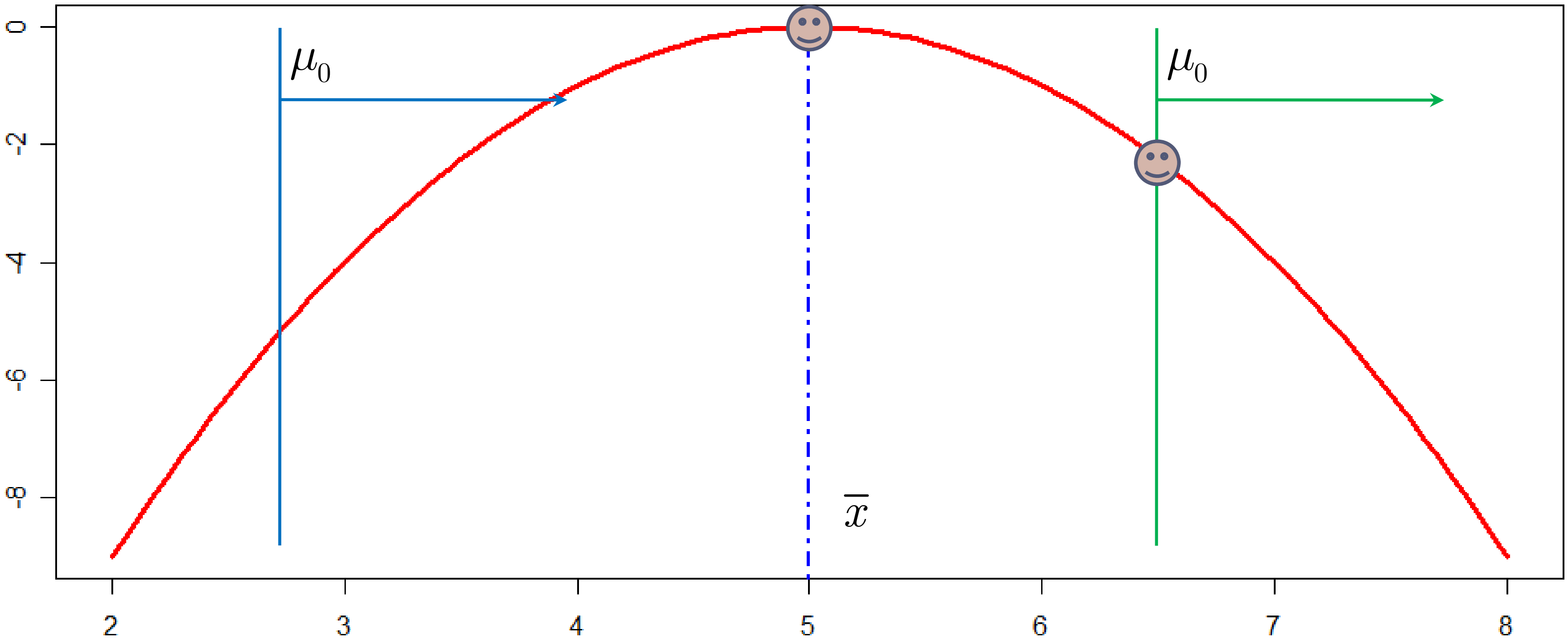
Certainly, when $\bar{x} \ge \mu_0$, $\bar{x}$ is still the MLE estimate of $\mu$.

However, when $\bar{x} < \mu_0$, the maximum is obtained at the boundary $\mu = \mu_0$.

Therefore,

$$\hat{\mu} = \begin{cases} \bar{x} & \text{if } \bar{x} \ge \mu_0 \\ \mu_0 & \text{if } \bar{x} < \mu_0 \end{cases}, \text{ and } L(\mu\,|\,\mathbf{x}) = \begin{cases} \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\sum_{i=1}^{n}(x_i-\bar{x})^2/(2\sigma^2)\right] & \text{if } \bar{x} \ge \mu_0 \\ \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\sum_{i=1}^{n}(x_i-\mu_0)^2/(2\sigma^2)\right] & \text{if } \bar{x} < \mu_0 \end{cases}$$

# Boundary conditions



# Normal MLE, both parameters

Let $X_1, \ldots, X_n$ be a sample from a normal population $\mathrm{N}(\mu, \sigma^2)$, where both $\mu$ and $\sigma^2$ are known. Then, the log likelihood function for $(\mu, \sigma^2)$ is

$$l(\mu, \sigma^2 | \mathbf{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\left(\sigma^2\right)^{-1}\sum_{i=1}^{n}(x_i - \mu)^2.$$

So we like to solve the optimization problem of

$$\max\ l(\mu, \sigma^2 | \mathbf{x}),$$

where $-\infty < \mu < \infty$ and $\sigma^2 > 0$. Let

$$\frac{\partial}{\partial\mu} l(\mu, \sigma^2 | \mathbf{x}) = \left(\sigma^2\right)^{-1}\sum_{i=1}^{n}(x_i - \mu) = \left(\sigma^2\right)^{-1}\left(\sum_{i=1}^{n} x_i - n\mu\right) = 0, \text{ and}$$

$$\frac{\partial}{\partial\sigma^2} l(\mu, \sigma^2 | \mathbf{x}) = -\frac{n}{2}\left(\sigma^2\right)^{-1} + \frac{1}{2}\left(\sigma^2\right)^{-2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0, \text{ and}$$

we have $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{n-1}{n}s^2.$

Hence, $\bar{X}$ and $\frac{n-1}{n}S^2$ is the MLE of $\mu$ and $\sigma^2$, respectively.

# Numeric solutions

In many cases, a maximum likelihood estimate is hard to obtain in a closed form, and we must resort to numeric solutions instead.

For example, let $X_1, \ldots, X_n$ be a random sample from a binomial $(k, p)$ population, where $p$ is known. The likelihood function is then

$$L(k \mid \mathbf{x}, p) = \prod_{i=1}^{n} \binom{k}{x_i} p^{x_i} (1-p)^{k-x_i}.$$

Since $k$ must be an integer, differentiation is difficult. However, the optimum $k$ must satisfy

$$\frac{L(k \mid \mathbf{x}, p)}{L(k-1 \mid \mathbf{x}, p)} \geq 1 \quad \text{and} \quad \frac{L(k \mid \mathbf{x}, p)}{L(k+1 \mid \mathbf{x}, p)} \geq 1$$

Therefore

$$(1-p)^n \geq \prod_{i=1}^{n} \left(1 - x_i / k\right) \quad \text{and} \quad (1-p)^n \leq \prod_{i=1}^{n} \left(1 - x_i / (k+1)\right).$$

Use numeric method can easily find the optimal $k$ when $k$ is not large.

# Invariance property of MLEs

***Invariance property of MLEs***

If $\hat{\theta}$ is the MLE of $\theta$, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

# Bernoulli MLE

Let $X_1, \ldots, X_n$ be a sample from a Bernoulli $(\theta)$ population, where $0 < \theta < 1$. Then, the likelihood function for $\theta$ is

$$L(\theta|\mathbf{x}) = \theta^{n_1}(1-\theta)^{n-n_1},\ n_1 = \sum_{i=1}^{n} X_i$$

and the log likelihood function is

$$l(\theta|\mathbf{x}) = \log L(\theta|\mathbf{x}) = n_1 \log\theta + (n - n_1)\log(1-\theta).$$

Maximize this function will yields

$$\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Therefore

$\bar{X}$ is the MLE of $\theta$.

Furthermore,

the MLE of $\sqrt{\theta(1-\theta)}$ is $\sqrt{\hat{\theta}(1-\hat{\theta})}$,

and the MLE of $\theta^2$ is $\hat{\theta}^2$.

# Making predictions

- What is the purpose of doing point estimation?
  - Estimate the parameters associated with a parametric distribution so that we can get full knowledge of the population
  - With the parameters estimated, we can calculate the value of the probability density (mass) for future values of the observation
  - In machine learning, we say density estimation
- How to calculate the probability density for new observations?

$$p(x^{\text{new}} \mid \theta^\star, \mathbf{x}) = p(x^{\text{new}} \mid \theta^\star)$$

$$p(\mathbf{x} \mid \theta) \Rightarrow \theta^\star = \arg\max p(\mathbf{x} \mid \theta) \Rightarrow p(x^{\text{new}} \mid \theta^\star, \mathbf{x}) = p(x^{\text{new}} \mid \theta^\star)$$

# Thank you very much