# Generalized likelihood ratio tests

The likelihood ratio test is optimal for simple vs. simple hypotheses.
Generalized likelihood ratio tests are for use when hypotheses are not simple.

They are not generally optimal, but are typically non-optimal in situations where no optimal test exists, and they usually perform reasonably well.

*Wide utility, plays the same role in testing as MLE's do in estimation.*

- Suppose that the observations $\boldsymbol{X} = (X_1, \ldots, X_n)$ have a joint distribution $f(\mathbf{x}|\theta)$.
- $H_0$ may specify that $\theta \in \omega_0$, where $\omega_0$ is a subset of the set of all possible $\theta$ values
- $H_1$ may specify that $\theta \in \omega_1$, where $\omega_1$ is disjoint from $\omega 0$.
- Let $\Omega = \omega_0 \cup \omega_1$.
- A plausible measure of the relative tenability of the hypotheses is the ratio of their likelihoods.
- Generalized likelihood ratio: discredit $H_0$ if small $\Lambda^* = \dfrac{\max\limits_{\theta \in \omega_0}[\text{lik}(\theta)]}{\max\limits_{\theta \in \omega_1}[\text{lik}(\theta)]}$
- In practical, use, instead, $\Lambda = \dfrac{\max\limits_{\theta \in \omega_0}[\text{lik}(\theta)]}{\max\limits_{\theta \in \Omega}[\text{lik}(\theta)]}$ and note $\Lambda = \min(\Lambda^*, 1)$
- Rejection region: small values.

$\Lambda \leq \lambda_0$. The threshhold $\lambda_0$ is chosen so that $P(\Lambda \leq \lambda_0 | H_0) = \alpha$,

desired significance level of the test.

## Generalized likelihood ratio tests

So far so good, but…

In order for the likelihood ratio test to have the significance level $\alpha$, $\lambda_0$ must be chosen so that $P(\Lambda \leq \lambda_0)=\alpha$ if $H_0$ is true.

- If the sampling distribution of $\Lambda$ under $H_0$ is known, we can determine $\lambda_0$.
- Generally, the sampling distribution is not of a simple form, what to do?

A theorem as the basis for an approximation to the null distribution:

> Under smoothness conditions on the probability density or frequency functions involved, the null distribution of $-2 \log \Lambda$ tends to a chi-square distribution with degrees of freedom equal to $\dim \Omega - \dim \omega_0$ as the sample size tends to infinity. ■

Dim $\Omega$ and dim $\omega_0$ are # of free parameters under $\Omega$ and $\omega_0$, respectively.

The above Gaussian example:

- $H_0$ completely specifies $\mu$ and $\sigma^2$, dim $\omega_0 = 0$
- Under $\Omega$, $\sigma$ is fixed, $\mu$ is free, dim $\Omega = 1$
- Null distribution of $-2 \log \Lambda$ is approximately (exactly!) $\chi_1^2$.

# Generalized likelihood Ratio Tests for the Multinomial Distribution

$H_0$ specifies that $p = p(\theta)$, $\theta \in \omega_0$, $\theta$ is a parameter (may be unknown);
$H_1$ cell probabilities are free except that they are nonnegative and sum to 1.
# of cells is $m$, $\Omega$ is the set consisting of m nonnegative numbers summing to 1.

Likelihood ratio's numerator is maximized when the MLE $\hat{\theta}$ is in place of $\theta$:

$$\max_{p \in \omega_0} \left( \frac{n!}{x_1! \cdots x_m!} \right) p_1(\theta)^{x_1} \cdots p_m(\theta)^{x_m} \rightarrow p_i(\hat{\theta})$$

Probabilities are unrestricted under $\Omega$, denominator is maximized by the unrestricted MLE's,

$$\hat{p}_i = \frac{x_i}{n}$$

The likelihood ratio is, therefore,

$$\Lambda = \frac{\frac{n!}{x_1! \cdots x_m!} p_1(\hat{\theta})^{x_1} \cdots p_m(\hat{\theta})^{x_m}}{\frac{n!}{x_1! \cdots x_m!} \hat{p}_1^{x_1} \cdots \hat{p}_m^{x_m}} = \prod_{i=1}^{m} \left( \frac{p_i(\hat{\theta})}{\hat{p}_i} \right)^{x_i}$$

Also, since $x_i = n\hat{p}_i$,

$$-2 \log \Lambda = -2n \sum_{i=1}^{m} \hat{p}_i \log \left( \frac{p_i(\hat{\theta})}{\hat{p}_i} \right) = 2 \sum_{i=1}^{m} O_i \log \left( \frac{O_i}{E_i} \right)$$

# Generalized likelihood Ratio Tests for the Multinomial Distribution

$$-2\log\Lambda = -2n\sum_{i=1}^{m}\hat{p}_i\log\left(\frac{p_i(\hat{\theta})}{\hat{p}_i}\right)$$
$$= 2\sum_{i=1}^{m}O_i\log\left(\frac{O_i}{E_i}\right)$$

where $O_i = n\hat{p}_i$ and $E_i = np_i(\hat{\theta})$ denote the observed and expected counts

- Under Ω, cell probabilities are free but summing to 1, so dim Ω=$m$−1
- Under $H_0$, probabilities $p_i(\hat{\theta})$ depend on a $k$-dimension parameter θ, dim $\omega_0$=$k$
- The large sample distribution of −2 log Λ is chi-square with dof=$m$−$k$−1.

**Pearson's chi-square statistic** is commonly used to test for goodness of fit

$$\chi^2 = \sum_{i=1}^{k}\frac{(O_i - E_i)^2}{E_i}$$

$$X^2 = \sum_{i=1}^{m}\frac{[x_i - np_i(\hat{\theta})]^2}{np_i(\hat{\theta})}$$

# Pearson's chi-square test

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad X^2 = \sum_{i=1}^{m} \frac{[x_i - np_i(\hat{\theta})]^2}{np_i(\hat{\theta})}$$

Likelihood ratio $$-2\log\Lambda = 2n\sum_{i=1}^{m} \hat{p}_i \log\left(\frac{\hat{p}_i}{p_i(\hat{\theta})}\right)$$

If $H_0$ is true, $n$ is large, $\hat{p}_i \approx p_i(\hat{\theta})$. Taylor series expansion of the function about $x_0$ is

$$f(x) = x\log\left(\frac{x}{x_0}\right) \qquad f(x) = (x - x_0) + \frac{1}{2}(x - x_0)^2\frac{1}{x_0} + \cdots$$

So,

$$-2\log\Lambda \approx 2n\sum_{i=1}^{m}[\hat{p}_i - p_i(\hat{\theta})] + n\sum_{i=1}^{m}\frac{[\hat{p}_i - p_i(\hat{\theta})]^2}{p_i(\hat{\theta})}$$

$2n\sum_{i=1}^{m}[\hat{p}_i - p_i(\hat{\theta})]$: 0, because $np$=1

$n\sum_{i=1}^{m}\frac{[\hat{p}_i - p_i(\hat{\theta})]^2}{p_i(\hat{\theta})}$: $\sum_{i=1}^{m}\frac{[x_i - np_i(\hat{\theta})]^2}{np_i(\hat{\theta})}$

Pearson's statistic & likelihood ratio are asymptotically equivalent under $H_0$.

But Pearson's chi-square test is easier in calculation and more widely used.

# Pearson's chi-square test

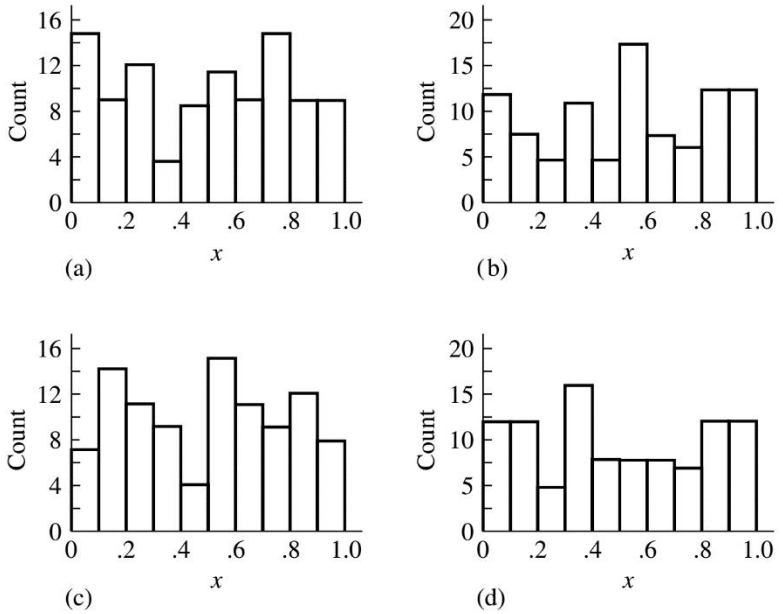
**Okay, but why do we choose multinomial distribution as $H_1$?**

Because in general, you cannot make it even more general/fuzzy/vague…

$H_0$: Any "real" distribution = all the parameters $p_i$ describing data in each bin are strongly correlated.

$H_1$: Multinomial distribution = almost no correlation between the parameters $p_i$ describing data in each bin, # of parameters = # of outcomes − 1. Way too many!

*E.g. Poisson: all $p_i$'s are determined by a single parameter.*

## Poisson Dispersion test

Likelihood ratio test, Pearson's chi-square test are w.r.t. the general form of $H_1$: cell probabilities are completely free (*that's why they are so useful*).
But if one has a specific $H_1$ in mind, better power may be obtained by testing against that $H_1$ rather than against a more general alternative.

*Poisson dispersion test is w.r.t. $H_0$: a distribution is Poisson.*

### Background and the need for it:

Key assumptions (the rate is constant, the counts in one interval of time or space are independent of the counts in disjoint intervals) are often not met.

- Count insects on leaves of plants. Different sizes, various locations on different plants; rate of infestation may well not be constant over the different locations.
- If insects hatched from eggs that were deposited in groups, then clustering of insects and the independence assumption might fail.
- Motor vehicle counts for traffic studies typically vary cyclically over time.

## Poisson Dispersion test

*Given counts* $x_1, \ldots, x_n$.

$H_0$: *Counts are Poisson with the common parameter* $\lambda$. $(\omega_0)$

$H_1$: *Counts are Poisson but with different rates,* $\lambda_1, \ldots, \lambda_n$. $(\Omega)$

MLE for $\omega_0$: $\hat{\lambda} = \overline{X}.$ MLE for $\Omega$: $\tilde{\lambda}_i = x_i$

$$\Lambda = \frac{\prod_{i=1}^{n} \hat{\lambda}^{x_i} e^{-\hat{\lambda}} / x_i!}{\prod_{i=1}^{n} \tilde{\lambda}_i^{x_i} e^{-\tilde{\lambda}_i} / x_i!} = \prod_{i=1}^{n} \left(\frac{\bar{x}}{x_i}\right)^{xi} e^{x_i - \bar{x}}$$

$$-2\log \Lambda = -2 \sum_{i=1}^{n} \left[ x_i \log\left(\frac{\bar{x}}{x_i}\right) + (x_i - \bar{x}) \right] = 2 \sum_{i=1}^{n} x_i \log\left(\frac{x_i}{\bar{x}}\right)$$

(Taylor expansion) $$\approx \frac{1}{\bar{x}} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \underline{n \hat{\sigma}^2 / \bar{x}}$$

D.o.f.=dim $\Omega$−dim $\omega_0$=$n$−1

Sensitive to (=has high power against) alternatives that are overdispersed relative to Poisson (e.g. negative binomial distribution.)

The ratio $\hat{\sigma}^2 / \bar{x}$ is sometimes used as a measure of clustering of data.

# Poisson Dispersion test

*Example.*

作为一个具体的例子，我们考虑美国国家科学与技术研究所的一项研究 (Steel 等 1980). 观测滤光片上石棉样纤维的个数，以制定石棉浓度的测量标准. 将石棉溶解在水中，然后散布在滤光片上，同时在滤光片上取直径 3 毫米的小孔，最后将其安放在透射电子显微镜下观察. 操作员计数下 23 个网格中的纤维数，得到的数据如下：

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 31 | 29 | 19 | 18 | 31 | 28 | 34 | 27 |
| 34 | 30 | 16 | 18 | 26 | 27 | 27 | 18 |
| 24 | 22 | 28 | 24 | 21 | 17 | 24 | |

Poisson dispersion test:

$$\frac{1}{\bar{x}}\sum(x_i-\bar{x})^2 = 26.56$$

Likelihood ratio test:

$$2\sum x_i \log\left(\frac{x_i}{\bar{x}}\right) = 27.11$$

D.o.f.=23−1=22, $p$-value=0.21

The evidence against the null hypothesis is not persuasive.

# Pearson's chi-square test: examples

*Bacterial clumps.* Let's test whether Poisson fits the data.

**例 9.5.2 (细菌凝块)** 在检验牛奶的细菌污染时，将 0.01 毫升的牛奶散放在 1 平方厘米的载玻片上，放在显微镜下观察，记录每个方格内的细菌凝块数. 乍一看，泊松模型可以非常合理地模拟凝块分布：据推测，凝块在牛奶中混合均匀，我们没有理由怀疑凝块束在一起. 然而，经过仔细检查，我们注意到两个可能的问题. 首先，受表面张力影响，奶滴下表面上的细菌可以粘附在与其相接触的玻璃载玻片上，导致这个胶片区域内的浓度增加. 其次，胶片的厚度不均匀，中心较薄，边缘较厚，引起细菌的浓度不均匀. 下表来自 Bliss 和 Fisher(1953)，汇总了 400 个方格上的凝块数.

| 每方格数 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 频　数 | 56 | 104 | 80 | 62 | 42 | 27 | 9 | 9 | 5 | 3 | 2 | 1 |

MLE gives us Poisson model with $\hat{\lambda} = \dfrac{0 \times 56 + 1 \times 104 + 2 \times 80 + \cdots + 19 \times 1}{400} = 2.44$

| 观测数 | 56 | 104 | 80 | 62 | 42 | 27 | 9 | 20 |
|---|---|---|---|---|---|---|---|---|
| 期望数 | 34.9 | 85.1 | 103.8 | 84.4 | 51.5 | 25.1 | 10.2 | 5.0 |
| $X^2$ 的成分 | 12.8 | 4.2 | 5.5 | 5.9 | 1.8 | 0.14 | 0.14 | 45.0 |

$H_0$: Poisson, $H_1$: multinomial

Chi-square statistic is $X^2$=75.4, dof=dim $\Omega$ (8−1)−dim $\omega_0(1)$ = 6.

Null hypothesis is conclusively rejected ($p$-value<0.005).

Reason: too many small counts and large counts to be Poisson.

## Poisson Dispersion test

*Example.* (细菌凝块)

$$\bar{x} = 2.44. \qquad \hat{\sigma}^2 = \frac{0^2 \times 56 + 1^2 \times 104 + \cdots + 19^2 \times 1}{400} - \bar{x}^2$$

$$= 4.59$$

the test statistic is

$$T = \frac{n\hat{\sigma}^2}{\bar{x}}$$

$$= \frac{400 \times 4.59}{2.40} = 752.7$$

Under $H_0$, $T \sim$ chi-square distribution with d.o.f.=399.
Chi-square with d.o.f. $m$ = sum of $m$ independent $N(0,\ 1)$ squared,
CLT says for large $m$ it is ~normal.

Chi-square: mean = d.o.f., variance = 2 d.o.f., $p$-value can be found by standardizing the statistic:

$$P(T \geqslant 752.7) = P\left(\frac{T - 399}{\sqrt{2 \times 399}} \geqslant \frac{752.7 - 399}{\sqrt{2 \times 399}}\right) \approx 1 - \Phi(12.5) \approx 0$$

Poission doesn't fit!

# Informal (graphical) techniques for assessing goodness of fit

# Hanging Rootograms

(Hanging rootograms) 使用直方图的方式显示观测值和拟合值的差. 为了解释悬挂根图及其构造，我们利用临床化学中的数据集 (Martin，Gudzinowicz 和 Fanger 1975). 下表给出了 152 个血清钾水平的实证分布. 在临床化学中，通常将这样的分布制作成包含“正态”值区间和病人化学水平的表格，以便比较确定其异常性. 表格中的分布经常用参数分布拟合，例如正态分布.

| Serum Potassium Levels | |
|---|---|
| Interval Midpoint | Frequency |
| 3.2 | 2 |
| 3.3 | 1 |
| 3.4 | 3 |
| 3.5 | 2 |
| 3.6 | 7 |
| 3.7 | 8 |
| 3.8 | 8 |
| 3.9 | 14 |
| 4.0 | 14 |
| 4.1 | 18 |
| 4.2 | 16 |
| 4.3 | 15 |
| 4.4 | 10 |
| 4.5 | 8 |
| 4.6 | 8 |
| 4.7 | 6 |
| 4.8 | 4 |
| 4.9 | 1 |
| 5.0 | 1 |
| 5.1 | 1 |
| 5.2 | 4 |
| 5.3 | 1 |

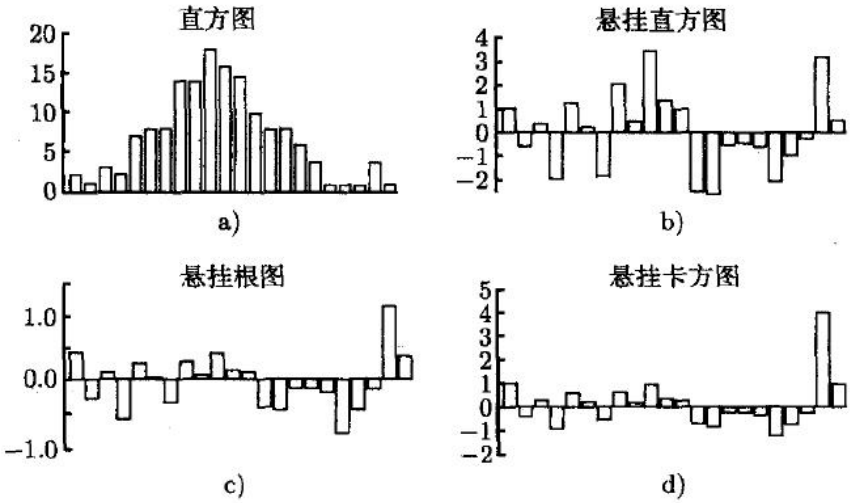



图 9.3 正态拟合血清钾数据的直方图、悬挂直方图、悬挂根图和悬挂卡方图

## Probability plots

Extremely useful graphical tool for *qualitatively* assessing the fit of data to a theoretical distribution.

Consider a sample of size $n$ from a uniform distribution on [0, 1]. Denote the *ordered* sample values by $X_{(1)}<X_{(2)}\cdots<X_{(n)}$ (called **order statistics**).

Question: What is the expectation of $X_{(j)}$?

**Make a guess!**

## Probability plots

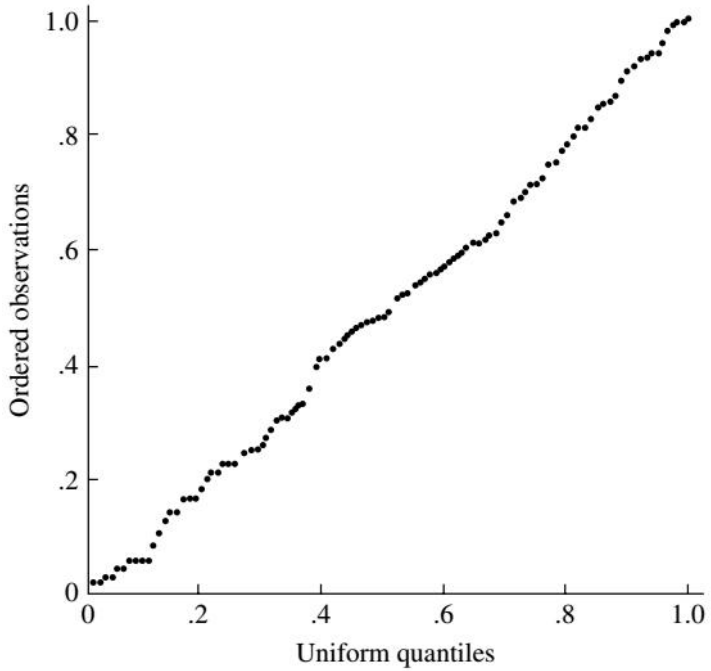
Extremely useful graphical tool for *qualitatively* assessing the fit of data to a theoretical distribution.

Consider a sample of size $n$ from a uniform distribution on [0, 1]. Denote the *ordered* sample values by $X_{(1)}<X_{(2)}\cdots<X_{(n)}$ (called **order statistics**).

The expectation of $X_{(j)}$ is

$$E(X_{(j)}) = \frac{j}{n+1}$$

Plot the ordered observations $X_{(1)},\ldots,X_{(n)}$ against their expected values $1/(n+1), \ldots, n/(n+1)$, in the case of a uniform distribution, it must be roughly linear.

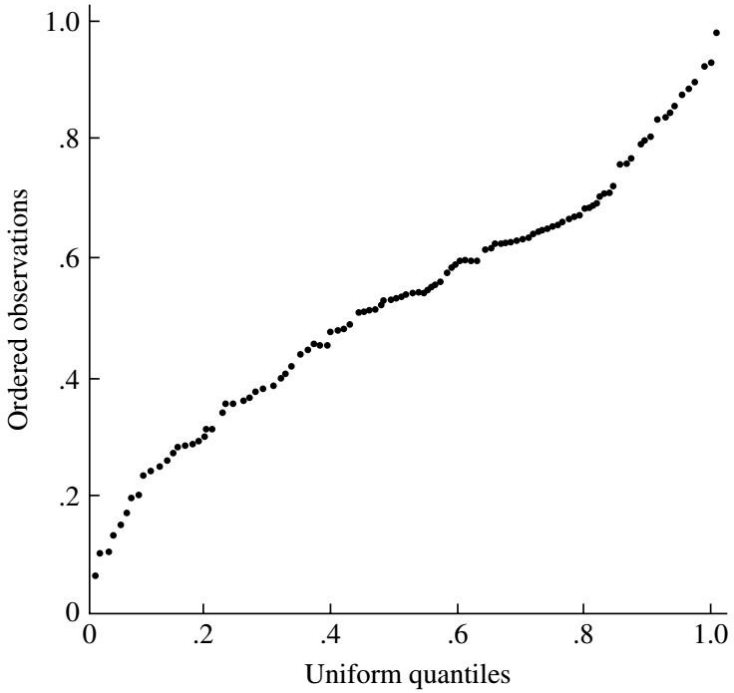


## Probability plots

Now suppose $Y=U_1/2+U_2/2$, then $Y$'s distribution is (by convolution) triangular:

$$f(y) = \begin{cases} 4y, & 0 \le y \le \frac{1}{2} \\ 4-4y, & \frac{1}{2} \le y \le 1 \end{cases}$$

Generate $Y_1$ to $Y_n$, plot $Y_{(1)}$ to $Y_{(n)}$ against the points $1/(n+1), \ldots, n/(n+1)$.




Data = triangular, theory = uniform, plot deviates from linearity:

- Left tail: order statistics > expected for a uniform distribution
- Right tail: opposite
- Tails are lighter than uniform

## Probability plots

The technique can be extended to other continuous probability laws this way:
If $X$ is a continuous random variable with a strictly increasing cdf, $F_X$, and if $Y = F_X(X)$, then $Y$ has a uniform distribution on [0, 1].

The transformation $Y = F_X(X)$ is known as the **probability integral transform**.

**Procedure**: Suppose $X$ is hypothesized to follow a certain distribution with cdf $F$. Given a sample $X_1, \ldots, X_n$, we plot

$$F(X_{(k)}) \quad \text{vs.} \quad \frac{k}{n+1}$$

or equivalently

$$X_{(k)} \quad \text{vs.} \quad F^{-1}\left(\frac{k}{n+1}\right)$$

In some cases, $F$ is of the form

$$F(x) = G\left(\frac{x-\mu}{\sigma}\right)$$

(e.g. normal)
μ=**location parameter**
σ=**scale parameter**

$$\frac{X_{(k)}-\mu}{\sigma} \quad \text{vs.} \quad G^{-1}\left(\frac{k}{n+1}\right)$$

Or the result would be approximately a straight line if the model were correct:

$$X_{(k)} \approx \sigma G^{-1}\left(\frac{k}{n+1}\right) + \mu$$

## Probability plots

Slight modifications of the procedure are sometimes used:

$$G^{-1}[k/(n+1)] \longrightarrow E(X_{(k)})$$

This is because it can be shown that

$$E(X_{(k)}) \approx F^{-1}\left(\frac{k}{n+1}\right)$$
$$= \sigma G^{-1}\left(\frac{k}{n+1}\right) + \mu$$

So we can plot $X_{(k)}$ vs. $E(X_{(k)})$ -- back to our first uniform example.

**Viewed from another perspective:**

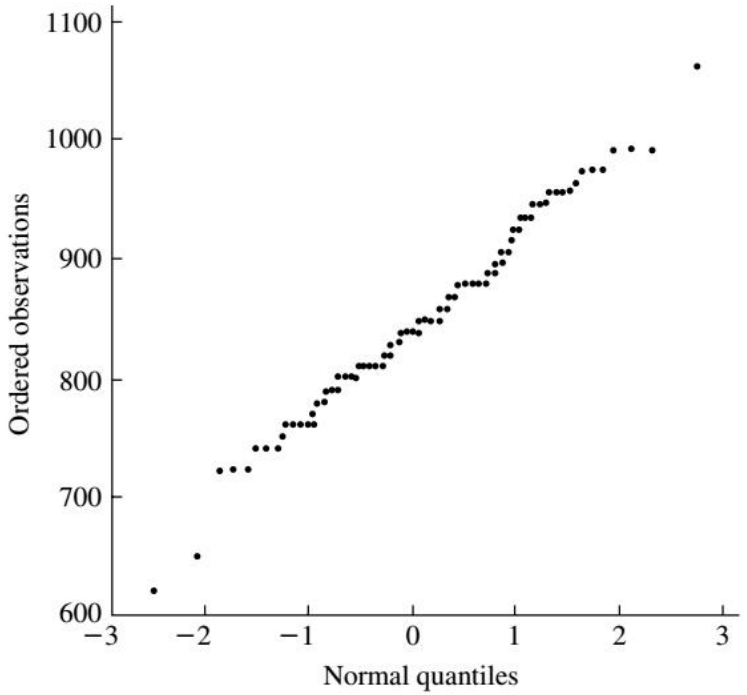
$F^{-1}[k/(n+1)]$ is the $k/(n+1)$ quantile of the distribution $F$, the point such that the probability that a random variable with cdf $F$ is less than it is $k/(n+1)$.

We are thus plotting the ordered observations (can be viewed as the observed or empirical quantiles) versus the quantiles of the theoretical distribution.

# Probability plots: examples

A set of 100 observations, which are Michelson's determinations of the velocity of light made from June 5, 1879 to July 2, 1879; 299,000 has been subtracted from the determinations to give the values listed [Stigler (1977)]:

| | | | | | |
|---|---|---|---|---|---|
| 850 | 960 | 880 | 890 | 890 | 740 |
| 940 | 880 | 810 | 840 | 900 | 960 |
| 880 | 810 | 780 | 1070 | 940 | 860 |
| 820 | 810 | 930 | 880 | 720 | 800 |
| 760 | 850 | 800 | 720 | 770 | 810 |
| 950 | 850 | 620 | 760 | 790 | 980 |
| 880 | 860 | 740 | 810 | 980 | 900 |
| 970 | 750 | 820 | 880 | 840 | 950 |
| 760 | 850 | 1000 | 830 | 880 | 910 |
| 870 | 980 | 790 | 910 | 920 | 870 |
| 930 | 810 | 850 | 890 | 810 | 650 |
| 880 | 870 | 860 | 740 | 760 | 880 |
| 840 | 880 | 810 | 810 | 830 | 840 |
| 720 | 940 | 1000 | 800 | 850 | 840 |
| 950 | 1000 | 790 | 840 | 850 | 800 |
| 960 | 760 | 840 | 850 | 810 | 960 |
| 800 | 840 | 780 | 870 | | |




Theoretical model=Gaussian, plot is close to straight line, a qualitatively good fit.

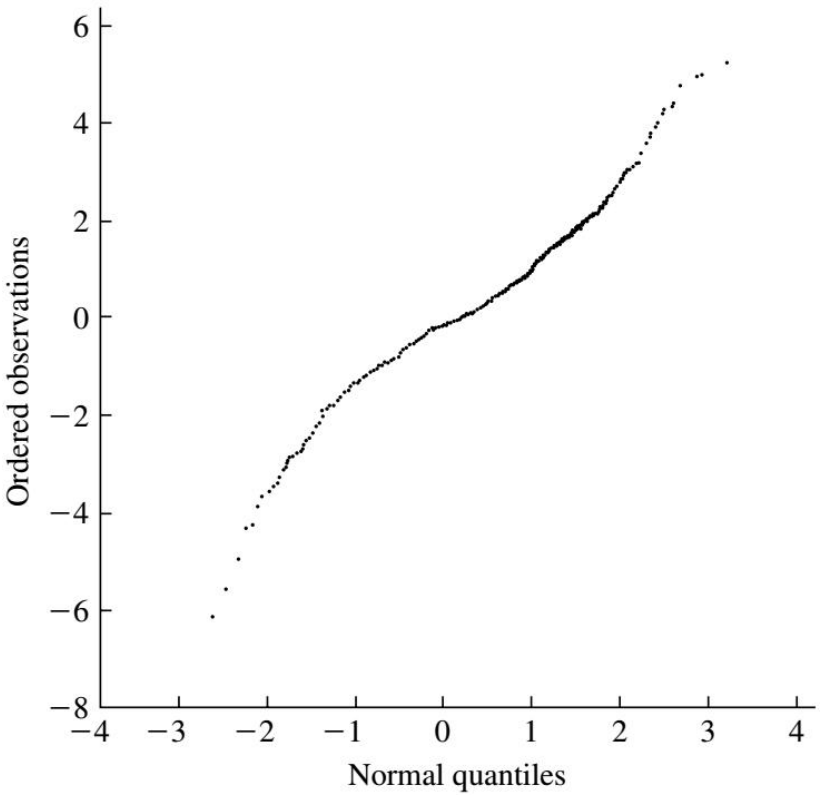
*Use cautions:*

- Probability plots are by nature monotonically increasing and they all tend to look fairly straight. Some experience is necessary in gauging "straightness."
- Simulations are very helpful in sharpening one's judgment

# Probability plots: examples

Nonnormal distributions. Below is a normal probability plot of 500 pseudorandom variables from a double exponential distribution:

$$f(x) = \tfrac{1}{2}e^{-|x|}, \qquad -\infty < x < \infty$$




Its tails die off at the rate exp(-|$x$|), slower than Gaussian, exp(-$x^2$).

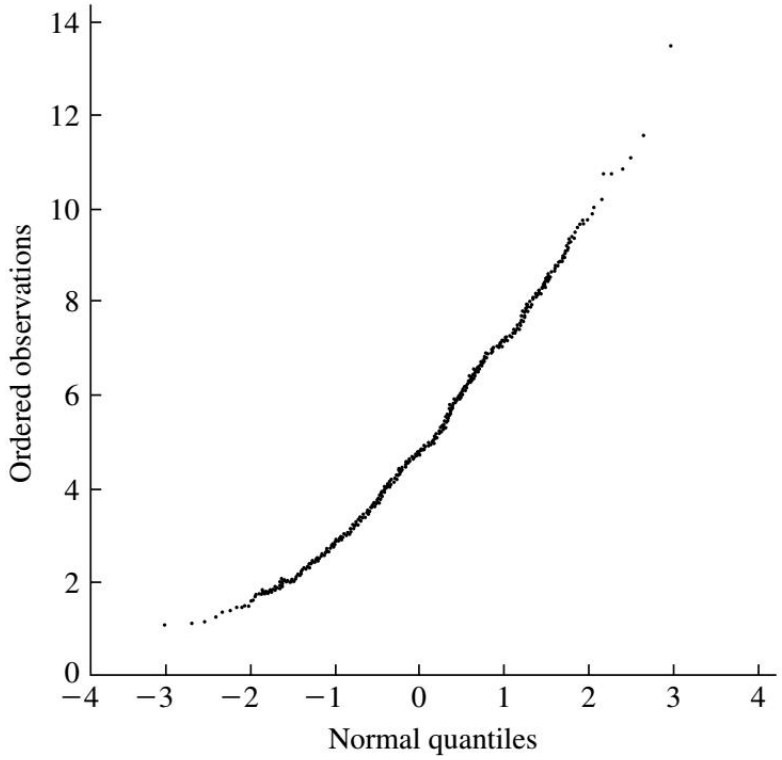
Plot bends down at the left and up at the right: observations in the left tail more negative than expected for Gaussian, observations in the right tail more positive.

The extreme observations were larger in magnitude than extreme observations from a Gaussian.

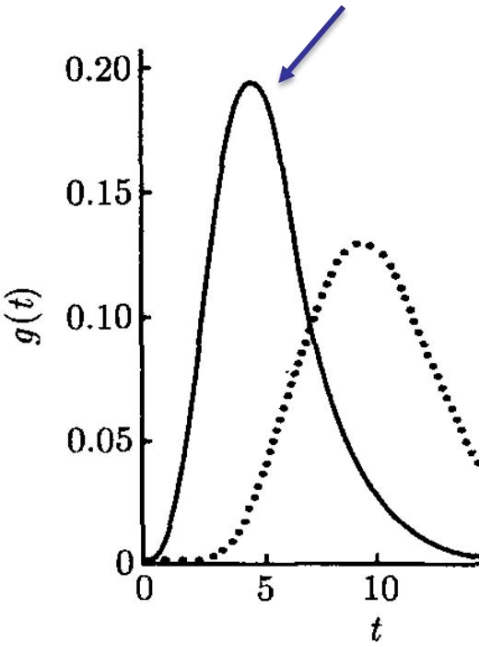
The tails of the double exponential are "heavier" than those of a Gaussian.

# Probability plots: examples

Nonnormal distributions: Gamma. 500 pseudorandom numbers from a gamma distribution with scale parameter $\lambda=1$, shape parameter $\alpha=5$.



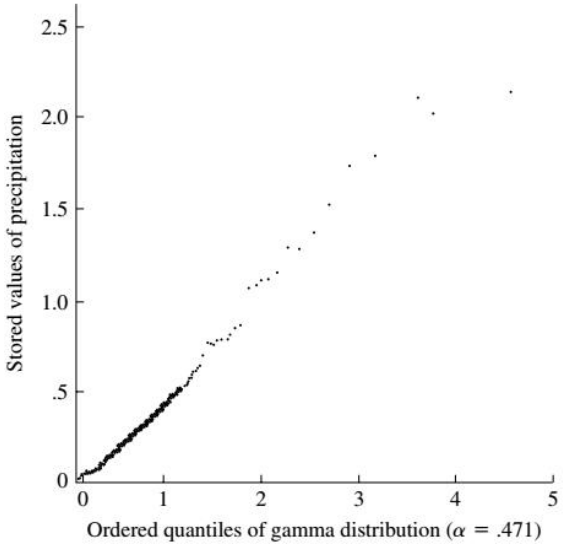

Bowlike – skewed/nonsymmetric.



# Probability plots: examples

Gamma probability plot of the precipitation amounts.

A computer was used to find the quantiles of a gamma distribution with parameter $\alpha = .471$ and $\lambda = 1$. The plot is observed sorted values of precipitation versus the quantiles.




Qualitatively, the fit is reasonable, there is no gross systematic deviation from a straight line.

**Probability plots for grouped data**（分组数据）. Suppose that the grouping gives the points $x_1, \ldots, x_{m+1}$ for the histogram's bin boundaries and that in the interval $[x_i, x_{i+1})$ there are $n_i$ counts, where $i = 1, \ldots, m$. We denote the cumulative frequencies by $N_j = \sum_{i=1}^{j} n_i$. Then $N_1 < N_2 < \cdots < N_m$ and $N_m =$ total size $n$. One can plot

$$x_{j+1} \qquad \text{vs.} \qquad G^{-1}\left(\frac{N_j}{n+1}\right), \qquad j = 1, \ldots, m$$

# Summarizing data

# Empirical cdf

Methods of describing and summarizing data that are in the form of one or more samples, or batches. These procedures often generate graphical displays, are useful in revealing the structure of data.

In the absence of a stochastic model, the methods are useful for purely descriptive purposes. If it is appropriate to entertain a stochastic model, the implications of that model for the method are of interest.

"Sample": $x_i$ are i.i.d. with some distribution function
*"Batch":* imply no such commitment to a stochastic model

Suppose $x_1, \ldots, x_n$ is a batch of numbers. The **empirical cdf (ecdf)** is defined as

$$F_n(x) = \frac{1}{n}(\#x_i \leq x)$$

Denote the ordered batch of numbers by $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.
Then if $x \leq x_{(1)}$, $F_n(x)=0$; if $x_{(1)} \leq x < x_{(2)}$, $F_n(x)=1/n$, if $x_{(k)} \leq x \leq x_{(k+1)}$, $F_n(x)=k/n \ldots$

Single observation with value $x$, $F_n$ has a jump of height $1/n$ at $x$;
$r$ observations with same value $x$, $F_n$ has a jump of height $r/n$ at $x$

$F(x)$ gives the probability that $X \leq x$
$F_n(x)$ gives the proportion of the collection of numbers less than or equal to $x$

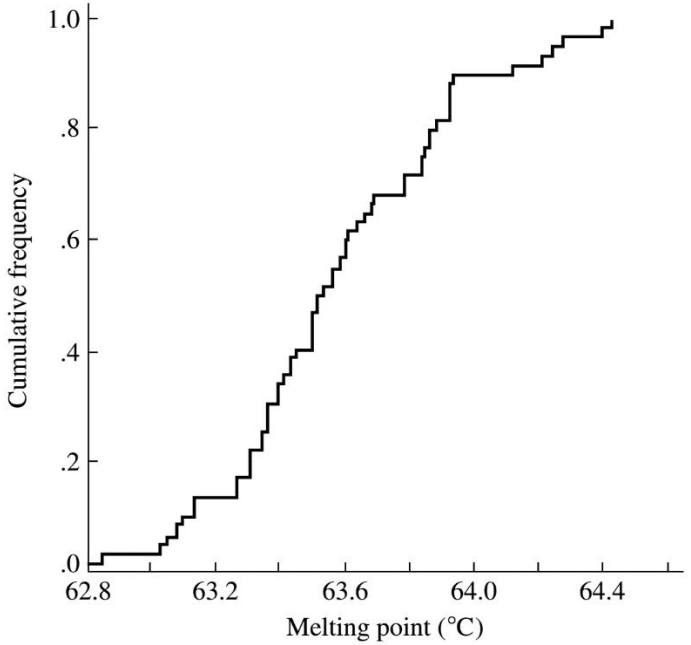
# Empirical cdf: example.

Example: White, Riethof & Kushnir (1960) 蜂蜡化学性质的研究数据，研究目的是通过一些化学试验，探测蜂蜡中人造蜡的存在（影响熔点）。作者得到59个纯蜂蜡的样本，熔点（摄氏度）如下：

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 63.78 | 63.45 | 63.58 | 63.08 | 63.40 | 64.42 | 63.27 | 63.10 |
| 63.34 | 63.50 | 63.83 | 63.63 | 63.27 | 63.30 | 63.83 | 63.50 |
| 63.36 | 63.86 | 63.34 | 63.92 | 63.88 | 63.36 | 63.36 | 63.51 |
| 63.51 | 63.84 | 64.27 | 63.50 | 63.56 | 63.39 | 63.78 | 63.92 |
| 63.92 | 63.56 | 63.43 | 64.21 | 64.24 | 64.12 | 63.92 | 63.53 |
| 63.50 | 63.30 | 63.86 | 63.93 | 63.43 | 64.40 | 63.61 | 63.03 |
| 63.68 | 63.13 | 63.41 | 63.60 | 63.13 | 63.69 | 63.05 | 62.85 |
| 63.31 | 63.66 | 63.60 | | | | | |

Conveniently summarizes the natural variability in melting points.

We see about 90% of the samples had melting points < 64.2 ◦ C, about 12% had melting points < 63.2 ◦ C.



## Empirical cdf: example.

Elementary statistical properties of the ecdf in the case in which $X_1, \ldots, X_n$ is a random sample from a continuous cdf, $F$:

It is convenient to express $F_n$ as:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i) \qquad I_{(-\infty,x]}(X_i) = \begin{cases} 1, & \text{if } X_i \le x \\ 0, & \text{if } X_i > x \end{cases}$$

In fact, the random variables $I_{(-\infty,x]}(X)$ are independent Bernoulli random variables:

$$I_{(-\infty,x]}(X_i) = \begin{cases} 1, & \text{with probability } F(x) \\ 0, & \text{with probability } 1 - F(x) \end{cases}$$

Thus, $nF_n(x)$ is a binomial random variable ($n$ trials, probability $F(x)$ of success)

$$E[F_n(x)] = F(x)$$

$$\text{Var}[F_n(x)] = \frac{1}{n}F(x)[1 - F(x)]$$

As an estimate of $F(x)$, $F_n(x)$ is unbiased and has a maximum variance at that value of $x$ such that $F(x) = 0.5$, that is, at the median. As $x$ becomes large or small, the variance tends to zero.

We considered $F_n$ for fixed $x$, but much deeper analysis focuses on the stochastic behavior of $F_n$ as a random function (consider all $x$ values simultaneously).

Surprisingly, it turns out that the following does not depend on $F$ if $F$ is continuous!

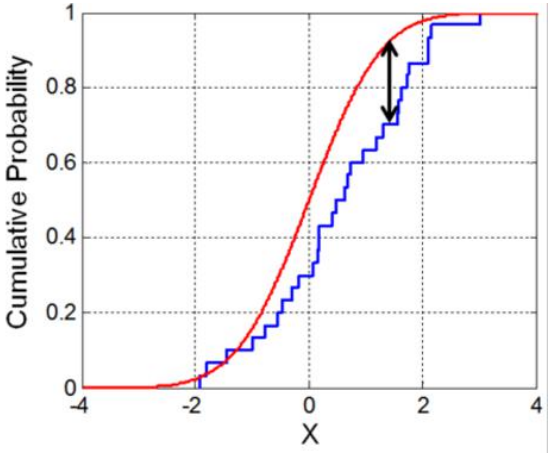
$$\max_{-\infty<x<\infty} |F_n(x) - F(x)|$$

This result forms the foundation of the famous and *extremely widely used* **Kolmogorov-Smirnov test**:

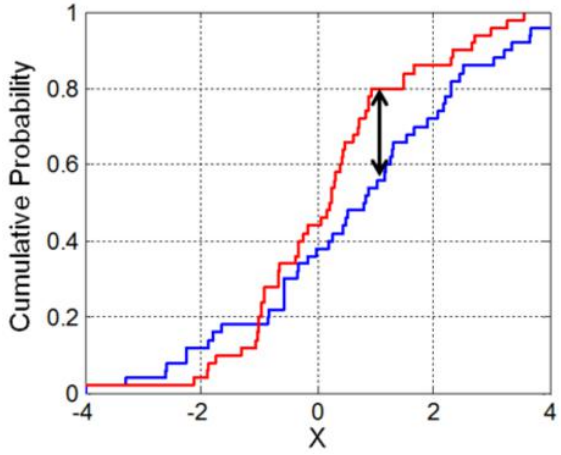
$$D_n = \sup_x |F_n(x) - F(x)|$$

This maximum difference corresponds to a unique $p$-value (ready from mathematician) for rejecting the hypothesis that the two (e)cdfs are from the same distribution.

1-sample K-S test: data vs. model




2-sample K-S test: data vs. data



**Quantile-quantile (Q-Q) plots** are useful for comparing distribution functions.
The $p$th quantile of the distribution was defined before as

$$F(x) = p \qquad x_p = F^{-1}(p)$$

In a Q-Q plot, the quantiles of one distribution vs. those of another.
$F(x)$: model for data of a control group; $G(y)$: model for data of a treatment group.

*Simplest effect #1*: increase expected response of every member of the treatment group by the same $h$. Then $y_p = x_p + h$, Q-Q plot is a straight line with slope 1, intercept $h$.
Cdf's relation:

$$G(y) = F(y - h)$$

*Simplest effect #2*: The response is multiplied by a constant $c$. Then the quantiles are related as $y_p = c\, x_p$ , Q-Q plot is a straight line with slope $c$, intercept 0. Cdf's relation

$$G(y) = F(y/c)$$

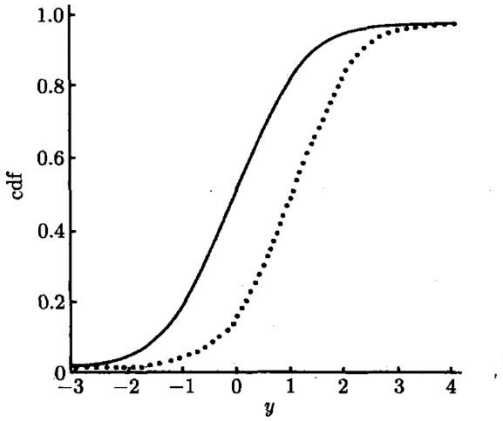



图 10.4 加法的试验效应. 实线是 $F(y)$，虚线是 $G(y) = F(y - h)$

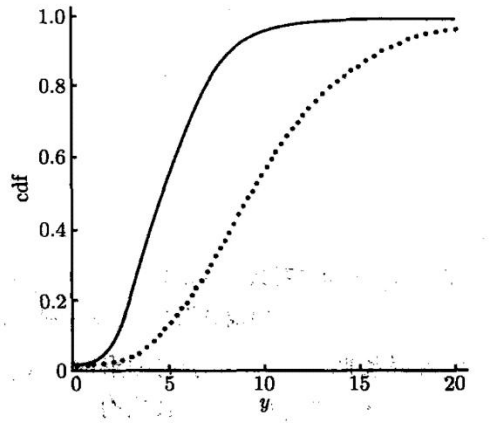



图 10.5 乘法的试验效应. 实线是 $F(y)$，虚线是 $G(y) = F(y/c)$

# Quantile-quantile plots

The effect of a treatment can be much more complicated

- an educational program that places very heavy emphasis on elementary skills may be a treatment that benefits weaker individuals but harm stronger individuals.

Given a batch of numbers, or a sample from a pdf, quantiles are constructed from the order statistics. Given $n$ observations and the order statistics $X_{(1)}, \ldots, X_{(n)}$, the $k/(n+1)$ quantile of data is assigned to $X_{(k)}$. → *we did this for probability plots to informally assess goodness of fit*.

To compare two batches of $n$ numbers with order statistics $X_{(1)}, \ldots, X_{(n)}$ and $Y_{(1)}, \ldots, Y_{(n)}$, a Q-Q plot is simply constructed by plotting the points $(X_{(i)}, Y_{(i)})$.

If the batches are of unequal size, an interpolation process can be used.

# Quantile-quantile plots

**例 10.2.3.1** Cleveland 等 (1974) 利用 Q-Q 图研究空气污染. 他们绘制了周日和平日各种变量值分布的分位数–分位数图 (图 10.6). 臭氧最大值的 Q-Q 图显示最高的分位数出现在平日，但其他的所有分位数都在周日较大. 对于一氧化碳、氮氧化物和气雾剂，分位数的差别随着浓度的增加而增大. 太阳辐射非常高和非常低的分位数在周日和平日近似相同 (大概相应于非常晴朗和乌云密布的日期)，但是对于中间的分位数，周日的分位数较大. ■

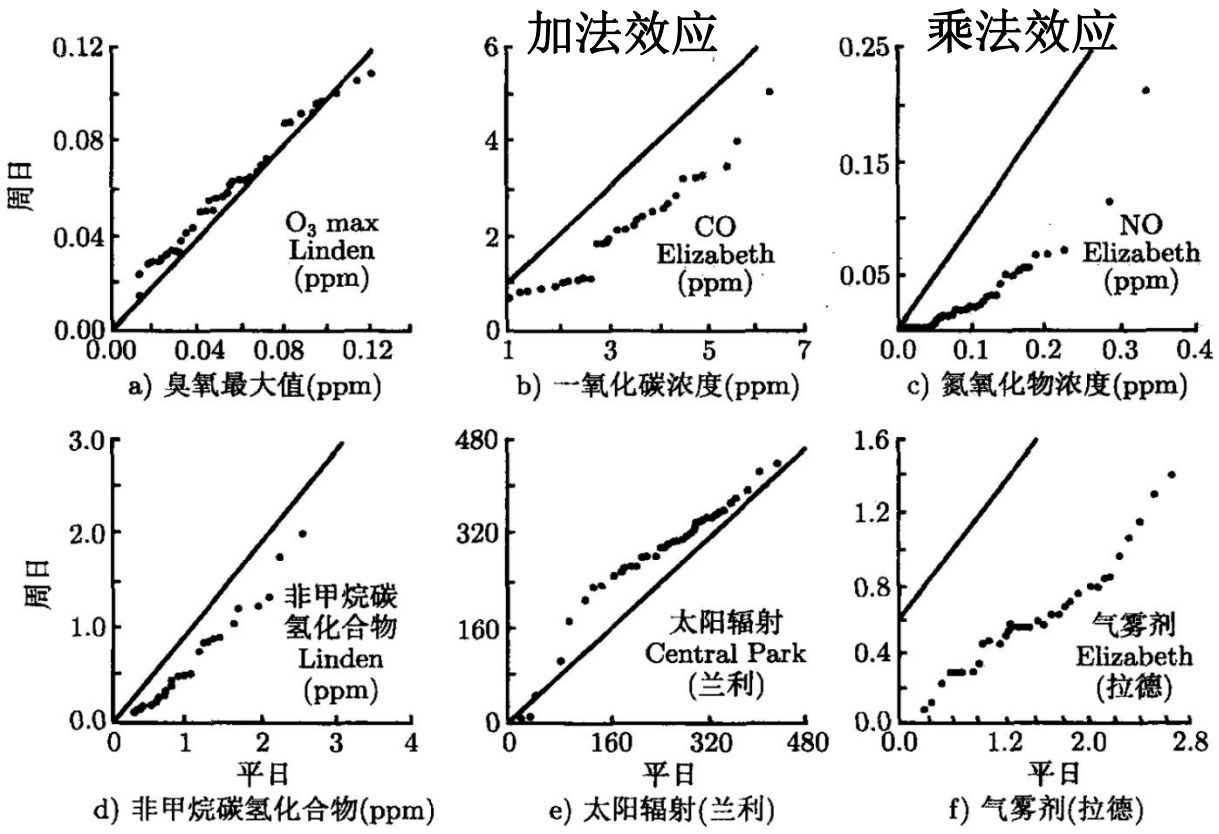



图 10.6 空气污染变量的 Q-Q 图

臭氧：最高分位值在平日，其他所有分位值周日较大

**CO, NO,** 气雾剂：分位数差别随浓度的增大而增大

太阳辐射：极高和极低的分位数在周日和平日近同

# Histograms, density curves, stem-and-leaf plots

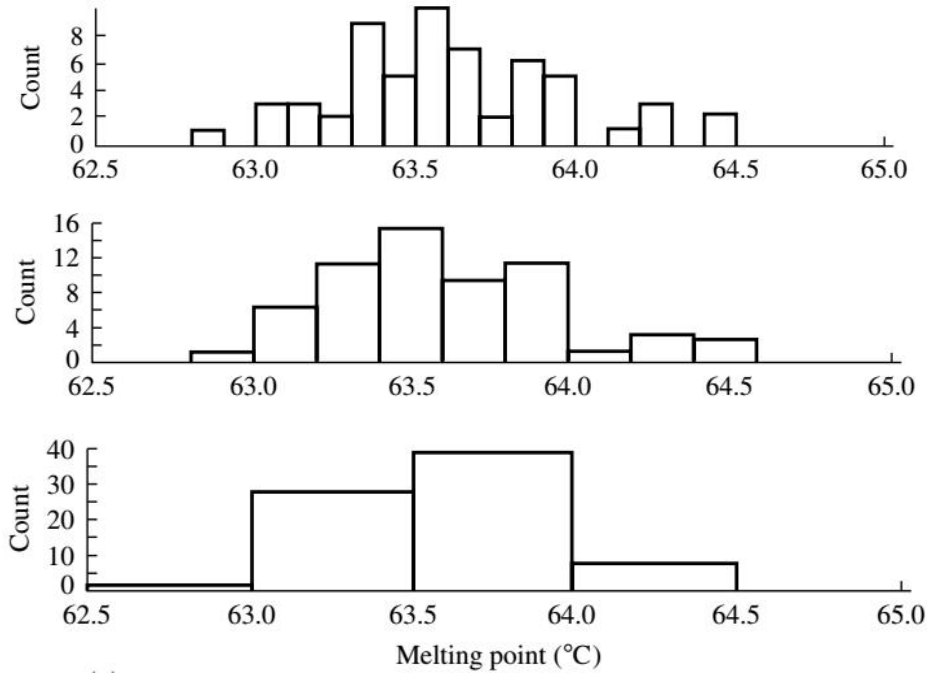
# Histograms

…displays the shape of the distribution of data values in the same sense that a density function displays probabilities.

- The range of data is divided into intervals/bins, plot counts OR proportion of the observations falling in each bin.
- Often recommended: plot the proportion of observations falling in the bin divided by the bin width, then the area under the histogram is 1.

Bin width:

- Too small: histogram too ragged
- Too wide: oversmoothed and obscured
- Usually made subjectively in an attempt to strike a balance…



# Histograms

- frequently used to display data for which there is no assumption of any stochastic model
- or, may be viewed as an estimate of the probability density: regarded in this sense, the histogram suffers from not being smooth…

*How to construct a smooth probability density estimate?*

Let $w(x)$ be a nonnegative, symmetric weight function, centered at zero, integrating to 1 (e.g. standard normal density). A rescaled version of $w$ is

$$w_h(x) = \frac{1}{h} w\left(\frac{x}{h}\right)$$

$h \to 0$, concentrated and peaked about zero. $h \to \infty$, spread out and flatter.
If $w$ is standard normal, $w_h(x)$ is normal density with standard deviation $h$.
If $X_1, \ldots, X_n$ is a sample from pdf $f$, an estimate of $f$ is

$$f_h(x) = \frac{1}{n} \sum_{i=1}^{n} w_h(x - X_i)$$

Called a **kernel probability density estimate**.

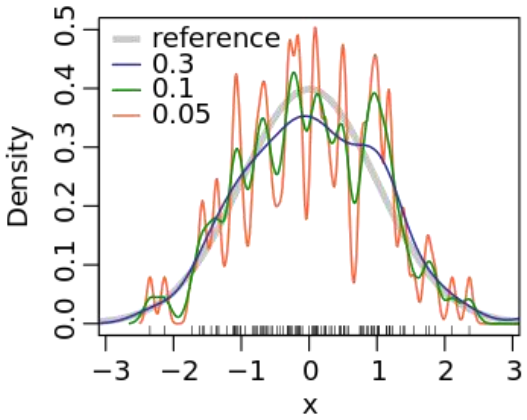
# Histograms

$$f_h(x) = \frac{1}{n}\sum_{i=1}^{n} w_h(x - X_i)$$

***"Smoothing kernel"***

A **kernel probability density estimate** consists of "hills" centered on the observations. If $w$ is standard normal, $w_h(x-X_i)$ is normal with parameters $X_i$, $h$

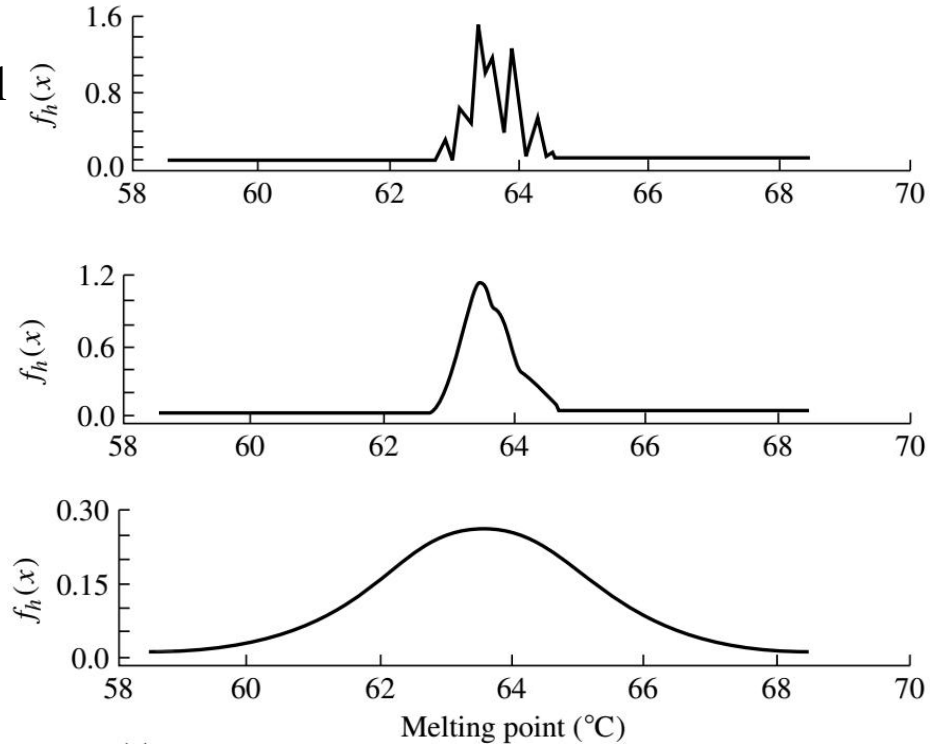
$h$ controls the **bandwidth** of the estimating function, controls its smoothness and corresponds to the bin width of the histogram.




$h$ too small

Good

$h$ too wide



# Stem-and-leaf plots

- One disadvantage of a histogram or a probability density estimate is that information is lost; neither allows the reconstruction of the original data.
- A histogram does not allow for calculating a statistic such as a median; one can tell which bin but not the actual value.

Tukey (1977): **stem-and-leaf plots** convey info about shape while retaining the numerical info. Easy for human AND computers.

列1：累计个数（至中位数，双向）

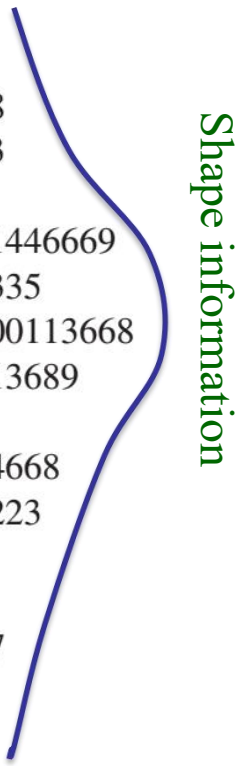
列2：每个茎的叶数

列3：数据X10的整数

列4：数据X10后的小数

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 63.78 | 63.45 | 63.58 | 63.08 | 63.40 | 64.42 | 63.27 | 63.10 |
| 63.34 | 63.50 | 63.83 | 63.63 | 63.27 | 63.30 | 63.83 | 63.50 |
| 63.36 | 63.86 | 63.34 | 63.92 | 63.88 | 63.36 | 63.36 | 63.51 |
| 63.51 | 63.84 | 64.27 | 63.50 | 63.56 | 63.39 | 63.78 | 63.92 |
| 63.92 | 63.56 | 63.43 | 64.21 | 64.24 | 64.12 | 63.92 | 63.53 |
| 63.50 | 63.30 | 63.86 | 63.93 | 63.43 | 64.40 | 63.61 | 63.03 |
| 63.68 | 63.13 | 63.41 | 63.60 | 63.13 | 63.69 | 63.05 | 62.85 |
| 63.31 | 63.66 | 63.60 | | | | | |

| | | STEM | LEAF |
|---|---|---|---|
| 1 | 1 | 628 | :5 |
| 1 | 0 | 629 | : |
| 4 | 3 | 630 | :358 |
| 7 | 3 | 631 | :033 |
| 9 | 2 | 632 | :77 |
| 18 | 9 | 633 | :001446669 |
| 23 | 5 | 634 | :01335 |
| | 10 | 635 | :0000113668 |
| 26 | 7 | 636 | :0013689 |
| 19 | 2 | 637 | :88 |
| 17 | 6 | 638 | :334668 |
| 11 | 5 | 639 | :22223 |
| 6 | 0 | 640 | : |
| 6 | 1 | 641 | :2 |
| 5 | 3 | 642 | :147 |
| 2 | 0 | 643 | : |
| 2 | 2 | 644 | :02 |



# Measures of location

# Measures of Location

Let us discuss simple numerical summaries of data that are useful when there is not enough data to justify constructing a histogram or an ecdf, or when a more concise summary is desired.

A **measure of location** is *a measure of the center of a batch of numbers.*

- If data result from different measurements of the same quantity, it is used in the hope that it is more accurate than any single measurement
- Otherwise, used as a simple summary of data

- The arithmetic mean

- The median

- The trimmed mean

- M estimates

## The arithmetic mean

The most commonly used measure of location

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example: A set of 26 measurements of the heat of sublimation of platinum (铂) from an experiment (Hampson & Walker, 1961).

| Heats of Sublimation of Platinum (kcal/mol) | | | | | | | |
|---|---|---|---|---|---|---|---|
| 136.3 | 136.6 | 135.8 | 135.4 | 134.7 | 135.0 | 134.1 | 143.3 |
| 147.8 | 148.8 | 134.8 | 135.2 | 134.9 | 146.5 | 141.2 | 135.4 |
| 134.8 | 135.8 | 135.0 | 133.7 | 134.4 | 134.9 | 134.8 | 134.5 |
| 134.3 | 135.2 | | | | | | |

A common statistical model for the variability of a measurement process is

$$X_i = \mu + \beta + \varepsilon_i$$

The $\varepsilon_i$ are usually assumed to be independent and identically distributed random variables with mean 0 and variance $\sigma^2$.

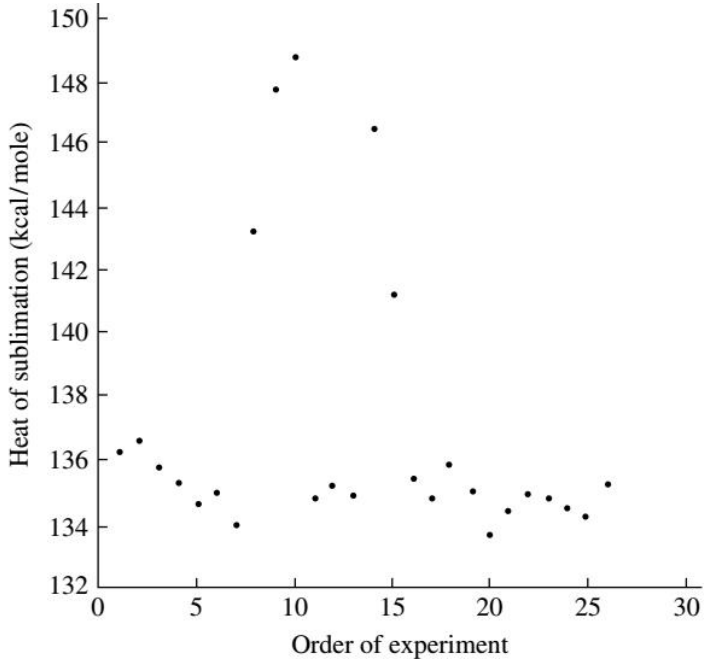
# The arithmetic mean

When observations are acquired sequentially, it is often informative to plot them in order: most striking aspect is the presence of five extreme observations that occurred in groups of three and two (**outliers**).

- improperly calibrated equipment
- Recording and transcription errors by equipment malfunctions
- Careful reexamination of data and circumstances
- But are often unexplainable aberrations

Is our model for measurement error appropriate for this data set?

The outliers occur in groups of 2 and 3, rather than being randomly scattered, making the independence model somewhat implausible.



# The arithmetic mean

Arithmetic mean is 137.05… look at the stem-and-leaf plot!
Clearly not a good descriptive measure of the "center" of this batch of numbers.

| | | |
|---|---|---|
| 1 | 1 | 133:7 |
| 4 | 3 | 134:134 |
| 11 | 7 | 134:5788899 |
| | 6 | 135:002244 |
| 9 | 2 | 135:88 |
| 7 | 1 | 136:3 |
| 6 | 1 | 136:6 |

High: 141.2 143.3 146.5 147.8 148.8

If the data are modeled as a sample from a probability law, an approximate $100(1-\alpha)\%$ confidence interval for the population mean can be obtained from CLT:

$$\bar{x} \pm z(\alpha/2)s_{\bar{x}}$$

*Blindly* applying it to data, with $\alpha=0.05$, we find (135.3, 138.8). Look at the plot!

By changing a single number, arithmetic mean can become arbitrarily large or small. Need careful attention!!

When data are automatically acquired, stored as files on disks, and not visually examined, this danger increases.

Need measures of location that are **robust**, insensitive to outliers.

# The median

Sample size is odd, median = middle value of the ordered observations
Sample size is even, median = average of the two middle values

Moving extreme observations does not affect the median at all! Very robust.
(above example: median=135.1)

When the data are a sample from a continuous probability law, the sample median can be viewed as an estimate of the population median, $\eta$, for which a simple confidence interval is of the form

$$(X_{(k)}, X_{(n-k+1)})$$

The coverage probability is

$$\begin{aligned} P(X_{(k)} \leq \eta \leq X_{(n-k+1)}) &= 1 - P(\eta < X_{(k)} \text{ or } \eta > X_{(n-k+1)}) \\ &= 1 - P(\eta < X_{(k)}) - P(\eta > X_{(n-k+1)}) \end{aligned}$$

$$P(\eta > X_{(n-k+1)}) = \sum_{j=0}^{k-1} P(\, j \text{ observations are greater than } \eta)$$

$$P(\eta < X_{(k)}) = \sum_{j=0}^{k-1} P(\, j \text{ observations are less than } \eta)$$

# The median

By definition, the median satisfies

$$P(X_i > \eta) = P(X_i < \eta) = \tfrac{1}{2}$$

We assume the $n$ observations to be i.i.d., distribution of # of observations > median is binomial, $n$ trials, ½ chance of success.

$$P(\text{exactly } j \text{ observations are greater than } \eta) = \frac{1}{2^n}\binom{n}{j}$$

$$P(\eta > X_{(n-k+1)}) = \sum_{j=0}^{k-1} P(\, j \text{ observations are greater than } \eta) = \frac{1}{2^n}\sum_{j=0}^{k-1}\binom{n}{j}$$

symmetry

$$P(X_{(k)} \le \eta \le X_{(n-k+1)}) = 1 - P(\eta < X_{(k)}) - P(\eta > X_{(n-k+1)})$$

$$= \quad 1 - \frac{1}{2^{n-1}}\sum_{j=0}^{k-1}\binom{n}{j}$$

## The median

These probabilities can be found from cdf of binomial distribution

$$\frac{1}{2^n}\sum_{j=0}^{k-1}\binom{n}{j} = P(Y \le k-1) \qquad Y\sim \text{Bin}\,(n, 1/2)$$

For example, with $n$=26:

| $k$ | $P(Y \le k)$ |
|---|---|
| 5 | .0012 |
| 6 | .0047 |
| 7 | .0145 |
| 8 | .0378 |
| 9 | .0843 |

If we choose $k$=8, $P(Y < k)= P(Y > n-k+1)= P(Y > 19)$=.0145, therefore $(X_{(8)}, X_{(19)})$ is a 1−.0145−.0145=97% confidence interval.

- This is an exact confidence interval, and does not depend on the form of the underlying cdf, only need continuous cdf and independent observations

- Platinum example: confidence interval is (134.8, 135.8) – based on mean: (135.3, 138.8). Much better.

# The trimmed mean (截尾均值)

**100α% Trimmed mean**

= arithmetic mean with lowest 100α% and highest 100α% discarded

α is generally recommended to be from 0.1 to 0.2.

Formally,

$$\bar{x}_\alpha = \frac{x_{([n\alpha]+1)} + \cdots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

- $[n\alpha]$ = the greatest integer $\leq n\alpha$
- Median = 50% trimmed mean

| | | | | | |
|---|---|---|---|---|---|
| | 1 | 1 | 133:7 | | |
| | 4 | 3 | 134:134 | | |
| | 11 | 7 | 134:5788899 | | |
| | | 6 | 135:002244 | | |
| | 9 | 2 | 135:88 | | |
| | 7 | 1 | 136:3 | | |
| | 6 | 1 | 136:6 | | |
| High: | 141.2 | 143.3 | 146.5 | 147.8 | 148.8 |

Platinum example:

- 20% trimmed mean = 135.9, rejecting 5 and 5 (0.2x26=5.2)
- Median = 135.1
- Arithmetic mean = 137.05

# Least squares estimate

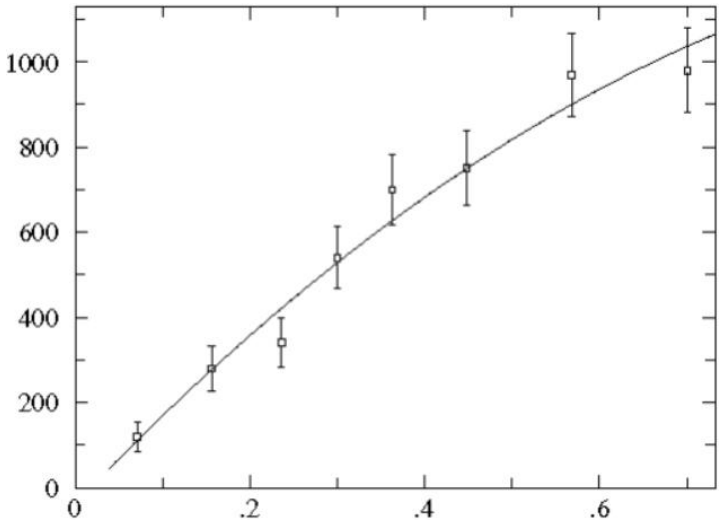
If underlying distribution is normal, sample mean = MLE of location parameter μ
Equivalently, it maximizes likelihood

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

→ minimize negative log likelihood (**least squares estimate**)

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$$

Each term ~ $\chi^2_1$, sum of $n$ terms ~ $\chi^2_{n-1}$, our bread-and-butter technique for curve fitting: **Minimizing $\chi^2$. A good fit should have reduced $\chi^2_n/n \sim 1$**. Why?



# Least squares estimate

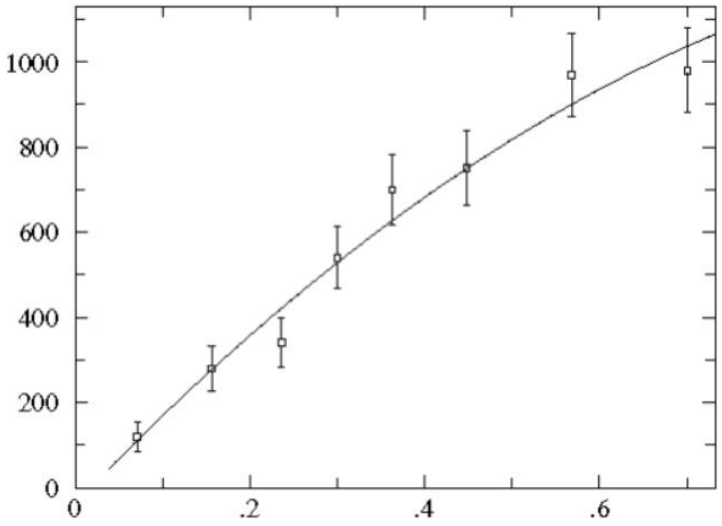
If underlying distribution is normal, sample mean = MLE of location parameter μ
Equivalently, it maximizes likelihood

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

→ minimize negative log likelihood (**least squares estimate**)

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$$

Each term ~ $\chi^2_1$, sum of $n$ terms ~ $\chi^2_{n-1}$, our bread-and-butter technique for curve fitting: **Minimizing $\chi^2$. A good fit should have reduced $\chi^2_n/n \sim 1$**. Why?




Deviation of data from model should be comparable to the length of the error bar (~1σ), so $\chi^2 \sim n$ (actually $n-1$), reduced $\chi^2 \sim 1$.

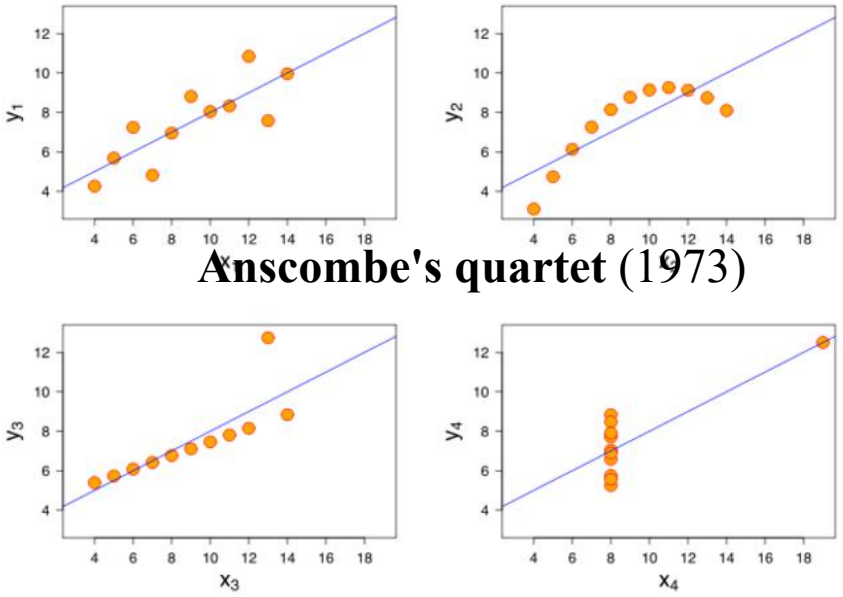
If reduced $\chi^2$ is $<<1$, often an indication that the measurement uncertainties are overestimated!

## Least squares estimate

$$\sum_{i=1}^{n} \left( \frac{X_i - \mu}{\sigma} \right)^2$$

**least squares**

Outliers have a great effect on this estimate, deviation is squared.

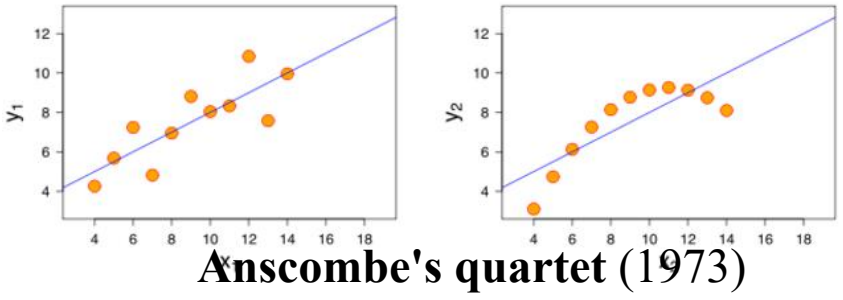



**Anscombe's quartet** (1973)

## Least squares vs. absolute deviations

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$$

**least squares**




**Anscombe's quartet** (1973)

Outliers have a great effect on this estimate, deviation is squared.

Median is the minimizer of $\sum_{i=1}^{n}\left|\frac{X_i-\mu}{\sigma}\right|$

**least absolute deviations**

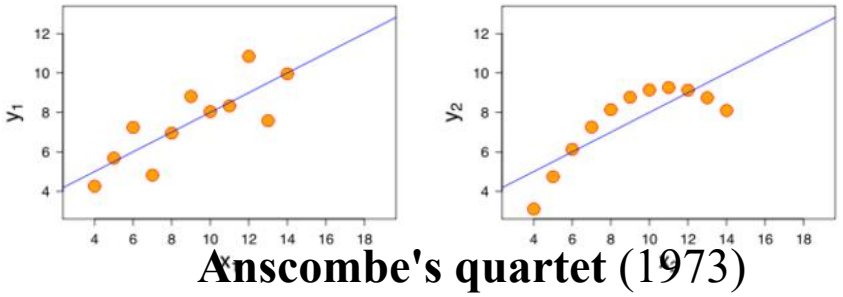
($d|x|/dx=\text{sign}(x)$, differentiating the sum yields $\sum_{i=1}^{n}\text{sign}\,(X_i-\mu)=0$ if $\mu$ is median)

| Least squares regression | Least absolute deviations regression |
|---|---|
| Not very robust | Robust |
| Stable solution | Unstable solution |
| Always one solution | Possibly multiple solutions |

## Least squares vs. absolute deviations

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$$

**least squares**




**Anscombe's quartet** (1973)

Outliers have a great effect on this estimate, deviation is squared.

Median is the minimizer of $\sum_{i=1}^{n}\left|\frac{X_i-\mu}{\sigma}\right|$

**least absolute deviations**

($\mathrm{d}|x|/\mathrm{d}x=\mathrm{sign}(x)$, differentiating the sum yields $\sum_{i=1}^{n}\mathrm{sign}\,(X_i-\mu)=0$ if μ is median)

### 最小二乘、一乘、零乘

最小零乘本质上是找一个数代表全体，跟它不一致的最少：众数（mode, 直方图的峰）

| Least squares regression | Least absolute deviations regression |
|---|---|
| Not very robust | Robust |
| Stable solution | Unstable solution |
| Always one solution | Possibly multiple solutions |

# M estimates

Huber (1981) proposed a class of **M estimates**, which are the minimizers of

$$\sum_{i=1}^{n} \Psi\left(\frac{X_i - \mu}{\sigma}\right)$$

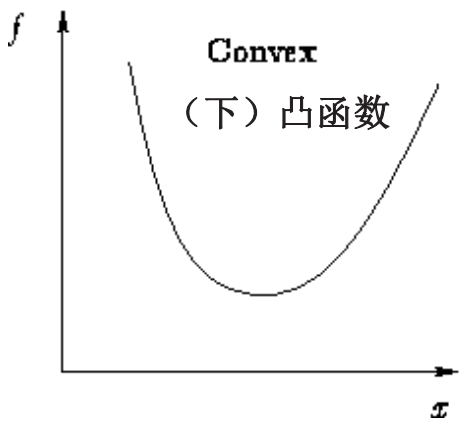
Ψ is a compromise between the weight functions for the mean and median.

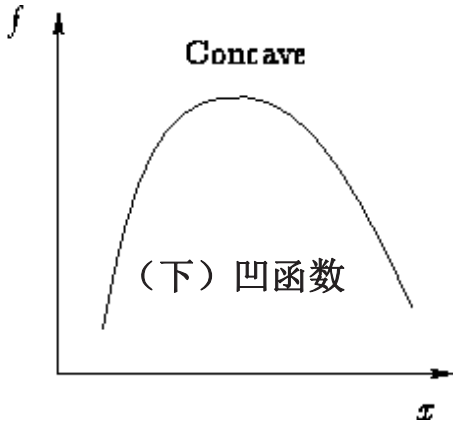
A wide variety of weight functions exist. E.g. by Huber himself:
weight functions that are quadratic near 0 and are linear beyond a cutoff point, $k$.

- $k \to \infty$: always quadratic = mean method
- $k \to 0$: always linear = median method
- A common choice is $k$=1.5, influence of observations >1.5σ away is reduced.
- If M is a **convex** function, minimizer is unique

Definition in your textbook is likely opposite to the international standard!!

Platinum example:
- 20% trimmed mean = 135.9
- Median = 135.1
- Arithmetic mean = 137.05
- M estimate ($k$=1.5)=135.38



## Comparison of location estimates

Which one is best? No simple answer… *Bear in mind what's being estimated by the location estimates and to what purpose the estimate is being put.*

Underlying distribution is symmetric:

- trimmed mean, sample mean, sample median, M all estimate center of symmetry.

When NOT symmetric, estimate 4 different population parameters:

- population mean/median/trimmed mean, a functional of the cdf determined by weight function $\Psi$.

## Comparison of location estimates

Which one is best? No simple answer… *Bear in mind what's being estimated by the location estimates and to what purpose the estimate is being put.*

Underlying distribution is symmetric:

- trimmed mean, sample mean, sample median, M all estimate center of symmetry.

When NOT symmetric, estimate 4 different population parameters:

- population mean/median/trimmed mean, a functional of the cdf determined by weight function $\Psi$.

Andrews et al. (1972): a large # of simulations from symmetric distributions

• 10% or 20% trimmed mean overall quite effective:

- Its variance never much larger than mean's (even for Gaussian, mean is optimal)
- can be a lot smaller when the underlying distribution is heavy-tailed relative to the Gaussian.

• Median is quite robust, but has a substantially larger variance in Gaussian case than the trimmed mean

• Trimmed mean and median have appealing simplicity, easy for statistics dummies

• M estimates perform quite well, generalize naturally to curve fitting etc. But hard to compute, have less immediate intuitive appeal.

• Often useful to compute more than one measure of location, compare them…

## Estimating variability of location estimates by bootstrap

Say $x_1,\ldots x_n$ is the realizations of i.i.d. random variables (cdf $F$), we need to investigate the variability and sampling distribution of a location estimate from a sample (size $n$).

Location estimate $\hat{\theta}$, we want to know its sampling distribution (determined by $n$, $F$).

(1) If we know $F$, $\hat{\theta}$ may be a complicated function of $x_1,\ldots x_n$, hard to calculate.

(2) And we actually don't know $F$.

**Way out for (1):**

- Suppose we knew $F$. We generate many ($B$ in #) samples of size $n$ from $F$; From each sample we calculate the value of $\hat{\theta}$.

- The empirical distribution of the resulting values $\theta_1^*,\ldots\theta_B^*$ is an approximation to the distribution function of $\hat{\theta}$.

- Calculate the standard deviation of $\theta_1^*,\ldots\theta_B^*$ approximating that of $\hat{\theta}$.

## Estimating variability of location estimates by bootstrap

Location estimate $\hat{\theta}$, we want to know its sampling distribution (determined by $n$, $F$).

(1) If we know $F$, $\hat{\theta}$ may be a complicated function of $x_1,\ldots x_n$, hard to calculate.

(2) And we actually don't know $F$.

**Way out for (2):**

- View the empirical cdf $F_n$ as an approximation to $F$, sample from $F_n$. But how to sample from $F_n$?

- $F_n$ is a discrete probability distribution that gives probability $1/n$ to each observed value $x_1,\ldots x_n$. We draw $B$ samples of size $n$ **with replacement** from the observed data, producing $\theta_1^*,\ldots\theta_B^*$. ("Draw $n$ values from $n$ values!!")

- Standard deviation of $\hat{\theta}$ is then estimated by

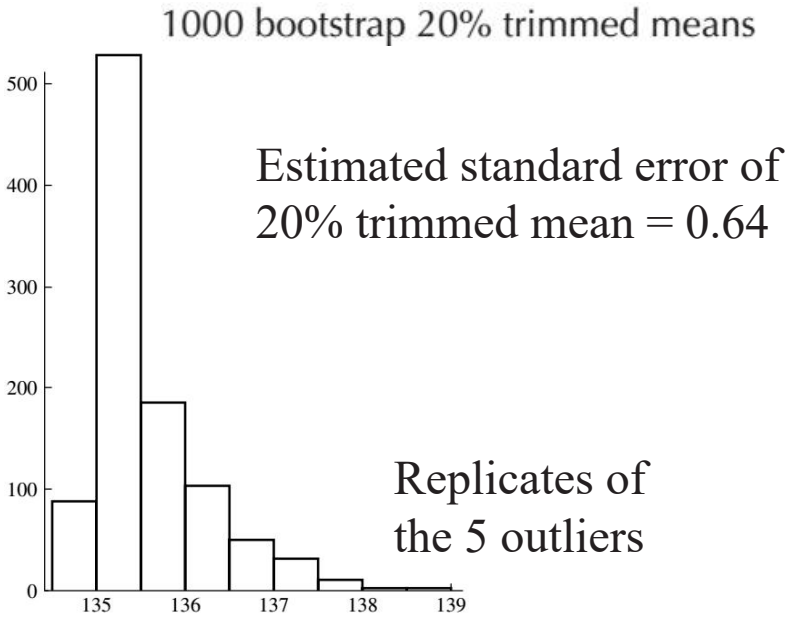
$$s_{\hat{\theta}} = \sqrt{\frac{1}{B}\sum_{i=1}^{B}(\theta_i^* - \bar{\theta}^*)^2}$$

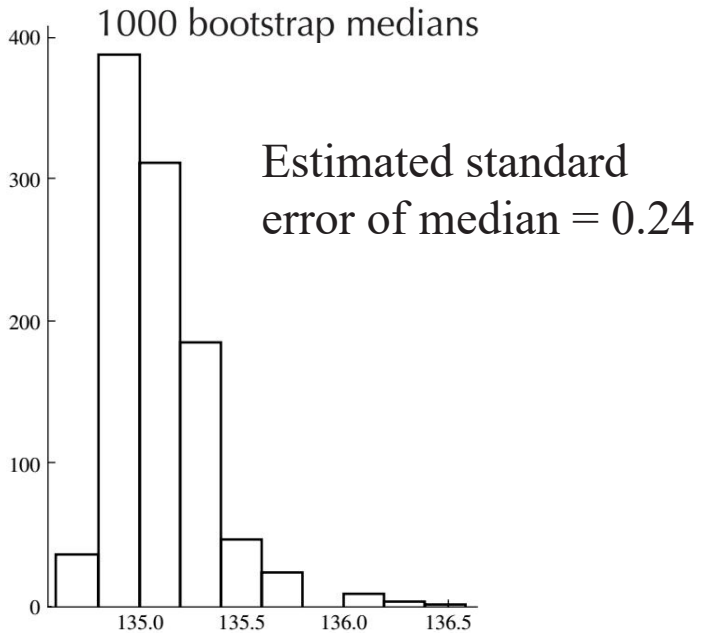
$\bar{\theta}^*$ is the mean of $\theta_1^*, \theta_2^*, \ldots, \theta_B^*$.

## Estimating variability of location estimates by bootstrap

*Example.* Platinum data again, using bootstrap to approximate sampling distribution of 20% trimmed mean and its standard error. 1000 samples of size $n$=26 drawn randomly from the collection of 26 values, with replacement.

The computer calculation tells us: 20% trimmed mean is not that robust, due to an extremely heavy-tailed distribution, a sample of 26 may contain many outliers.




Accuracy of bootstrap estimates?

- The accuracy of $F_n$ as an estimate of $F$
- dependence of the distribution of the statistic $\hat{\theta}$ on $F$ (sensitive→sample size large)

# Measures of dispersion

## Measure of dispersion

… gives a numerical indication of "scatteredness" of a batch of numbers.
Simple summaries of data often = measure of location + measure of dispersion

Most commonly used is sample standard deviation $s$,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

(Q: is $s$ an unbiased estimate of σ?)

If the observations are a sample from Gaussian with variance $\sigma^2$ (we proved it!),

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

→ confidence intervals for $\sigma^2$, but not robust against deviations from normality.

Sample standard deviation is sensitive to outliers. Simple robust measures:

**(1) Interquartile range** (IQR, 四分位差)
- difference between two sample quantiles ($25^{th}$ and $75^{th}$ percentiles)

**(2) Median absolute deviation from the median** (MAD, 中位数绝对偏差)
- median of the numbers $|x_i - \tilde{x}|$.

Sample median

## Measure of dispersion

**(1) Interquartile range** (IQR, 四分位差) -
difference between two sample quantiles (25th and 75th percentiles)

**(2) Median absolute deviation from the median** (MAD, 中位数绝对偏差)
- median of the numbers $|x_i - \tilde{x}|$.

Sample median

For Gaussian, IQR, MAD are converted into σ estimates by dividing by 1.35, 0.675.
(Q: how to do this conversion?)

Now compare all three measures of dispersion for the platinum data:

$s = 4.45$ - heavily influenced by outliers

$\frac{IQR}{1.35} = 1.26$

$\frac{MAD}{.675} = .934$

give measures of spread of central portion of data

| | | |
|---|---|---|
| 1 | 1 | 133:7 |
| 4 | 3 | 134:134 |
| 11 | 7 | 134:5788899 |
| | 6 | 135:002244 |
| 9 | 2 | 135:88 |
| 7 | 1 | 136:3 |
| 6 | 1 | 136:6 |

High: 141.2 143.3 146.5 147.8 148.8

# Boxplots

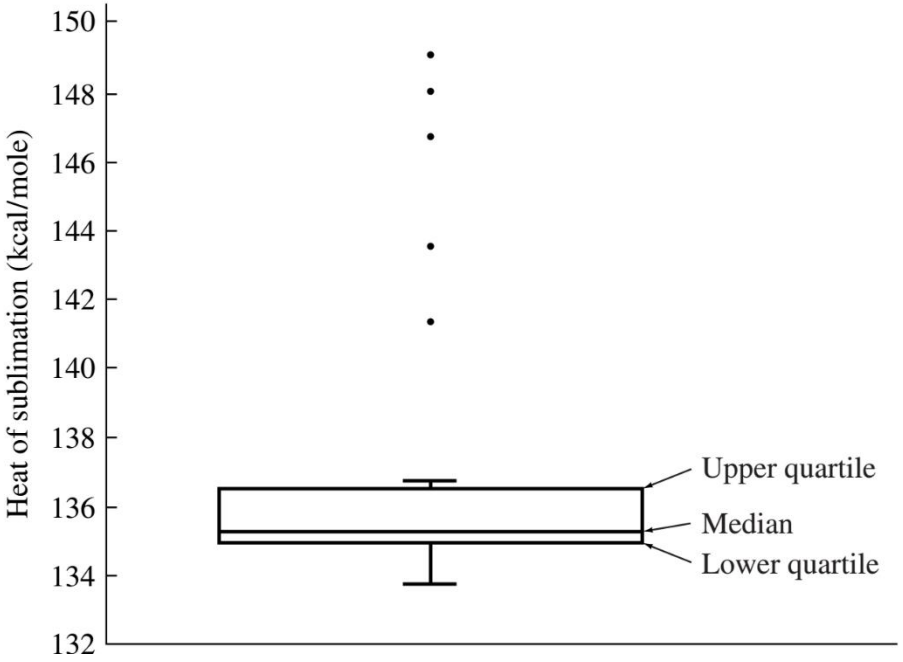
## Boxplots

Invented by Tukey, showing

- A measure of location (the median)
- A measure of dispersion (the interquartile range, IQR)
- Presence of possible outliers
- Indication of symmetry or skewness

Construction procedure:

- Horizontal lines at median, upper & lower quartiles
- Make it a box
- A vertical line from upper quartile to the most extreme data point that is within a distance of 1.5 (IQR) of the upper quartile.
- Same for lower quartile, add hats
- Data points beyond the ends are marked with • or *

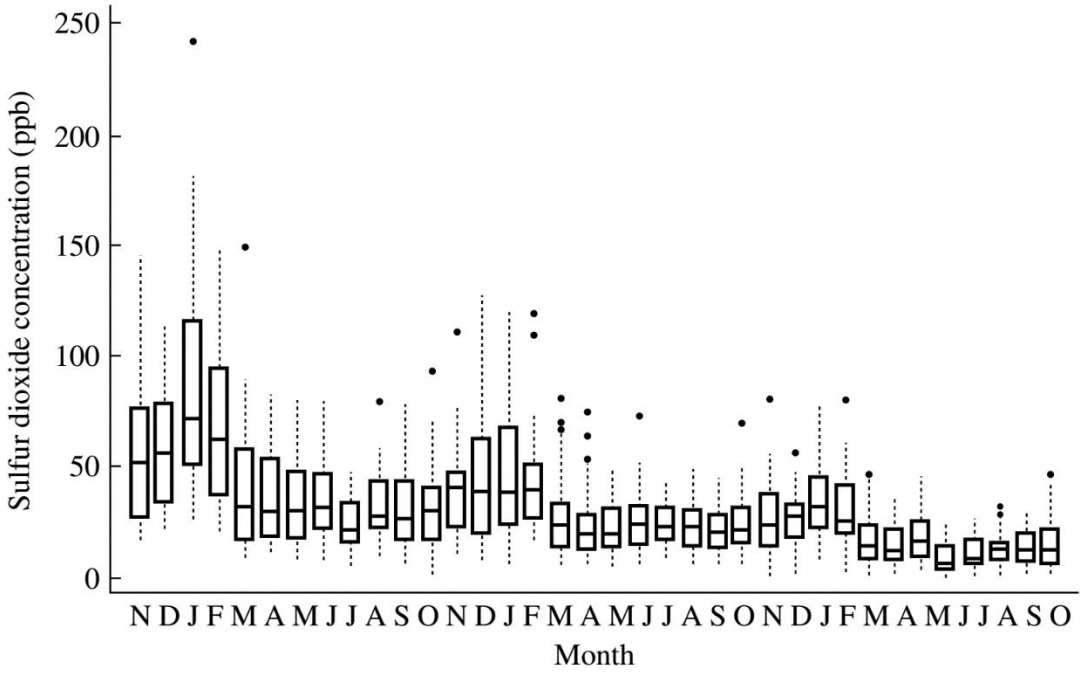



Indicating that central part of distribution is skewed toward high values.

## Boxplots

Chambers et al. (1983): The data plotted are daily maximum concentrations in parts per billion of sulfur dioxide in Bayonne, N.J., from Nov 1969 to Oct 1972 grouped by month. There are thus 36 batches, each of size about 30.

- A general reduction in $SO_2$ through time due to gradual conversion to low Sulphur fuels
- Higher concentrations during winter months due to using heating oil
- Skewed toward high values
- Spread is larger when general level of concentration is higher




*Very effective method of presenting and summarizing data, generally useful for comparing batches of numbers.*

# Exploring relationships with scatterplots

## Linear vs. logarithmic plots

Allison and Cicchetti (1976) examined the relationships of possible correlates of sleep behavior in mammals.

小短尾鼩鼱（qú jīng)

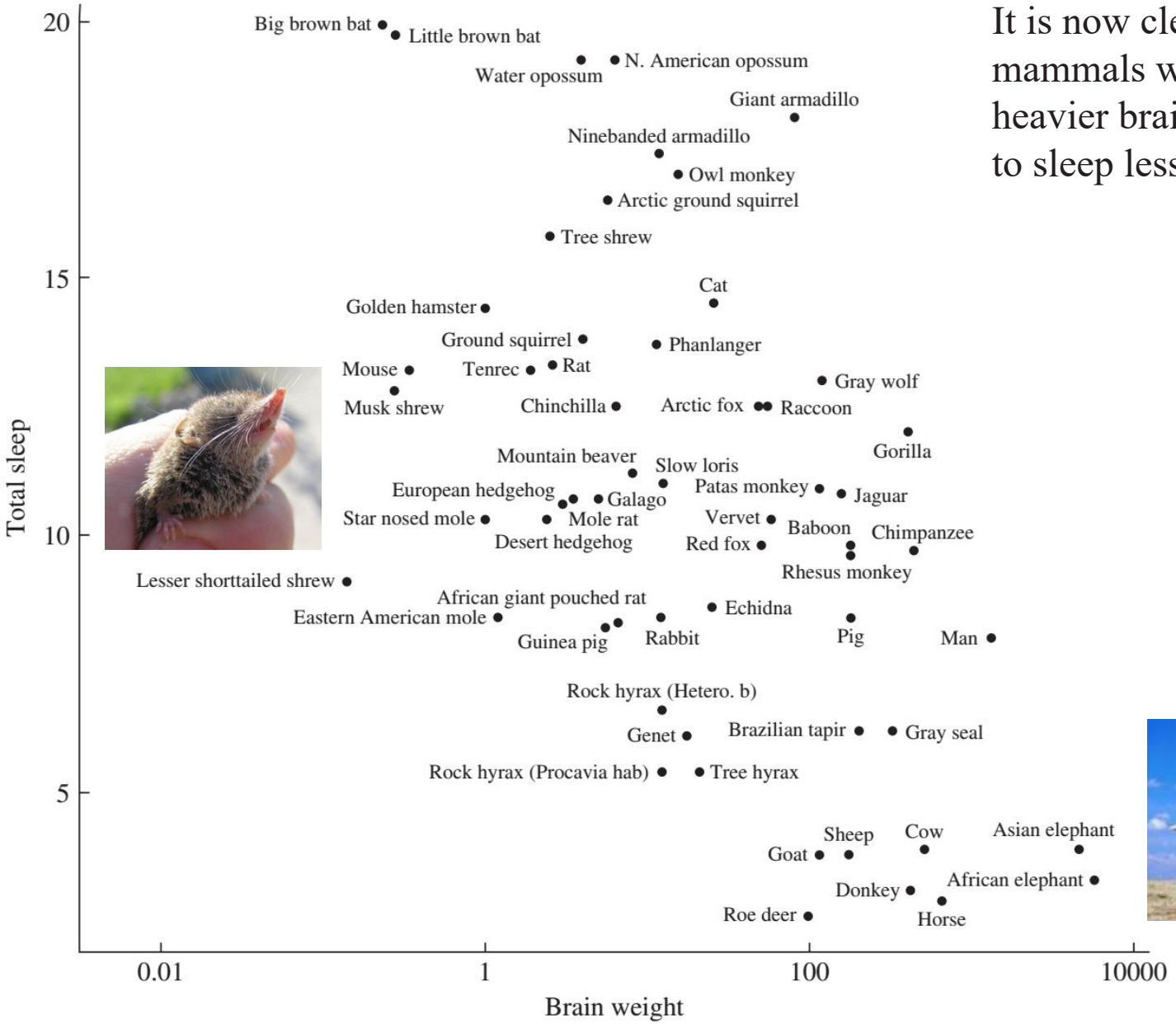
Two mammals with very large brains sleep very little, otherwise no relationships are apparent.

There is in fact a relationship, obscured because brain weights vary over orders of magnitude:

- 0.14g (lesser short-tailed shrew)
- 5,712g (African elephants)

Much more informative to plot sleep vs. the logarithm of brain weight

# Linear vs. logarithmic plots




It is now clear that mammals with heavier brains tend to sleep less!



## Scatterplots

Correlation coefficients: simple numerical summary of the strength of a relationship.

**Pearson correlation coefficient**:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

*r* measures the strength of a ***linear*** relationship.

- brain weight vs. sleep: −0.36; log brain weight vs. sleep: −0.56.
- different because a nonlinear transformation is applied and *r* measures the strength of a linear relationship.

**Rank correlation coefficient** (秩相关系数):

- brain weights are replaced by their ordered ranks (1, 2, . . .)
- sleeping times are replaced by their ranks
- Pearson correlation coefficient of the pairs of ranks is computed (− 0.39)

Advantages:

- insensitive to outliers
- invariant under any monotone transformation (same for log or not).