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 σ^2 (we proved it!),

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

→ confidence intervals for σ^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

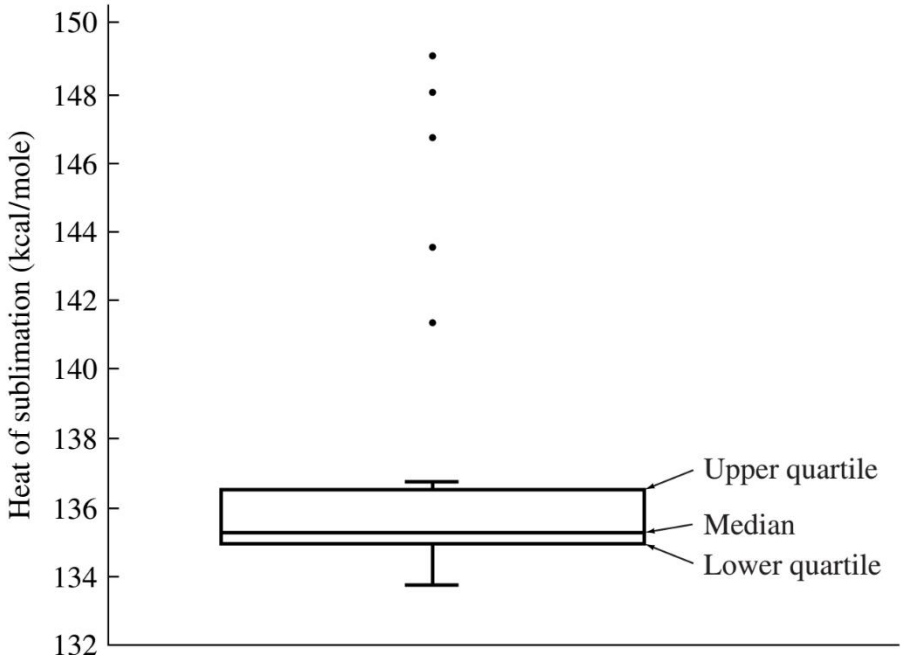
High: 141.2 143.3 146.5 147.8 148.8

Boxplots

Boxplots

Invented by Tukey, showing

- A measure of location (the median)
- A measure of dispersion (the interquartile range, IQR, |25th - 75th percentiles|)
- Presence of possible outliers
- Indication of symmetry or skewness



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

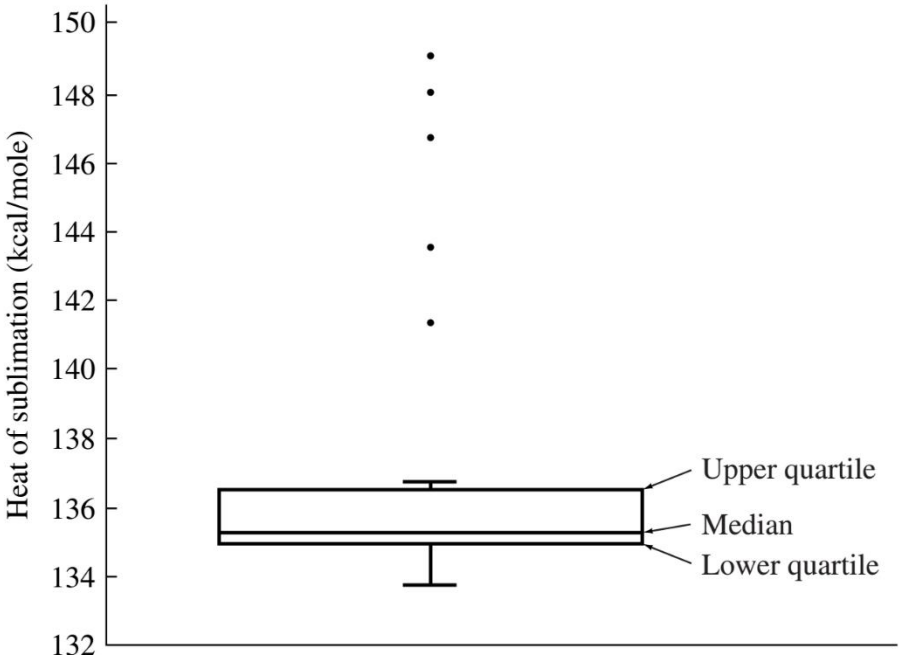
Boxplots

Invented by Tukey, showing

- A measure of location (the median)
- A measure of dispersion (the interquartile range, IQR, |25th - 75th percentiles|)
- 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 *

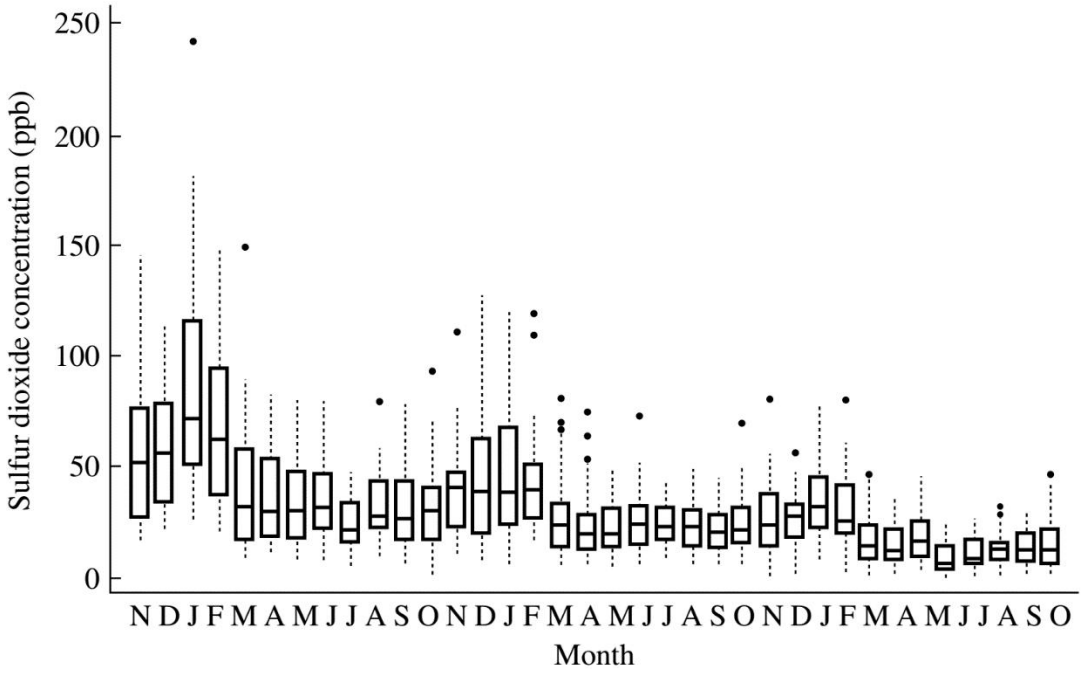


Indicating that central part of distribution is skewed toward high values.

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 SO2 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 - **continued**

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)

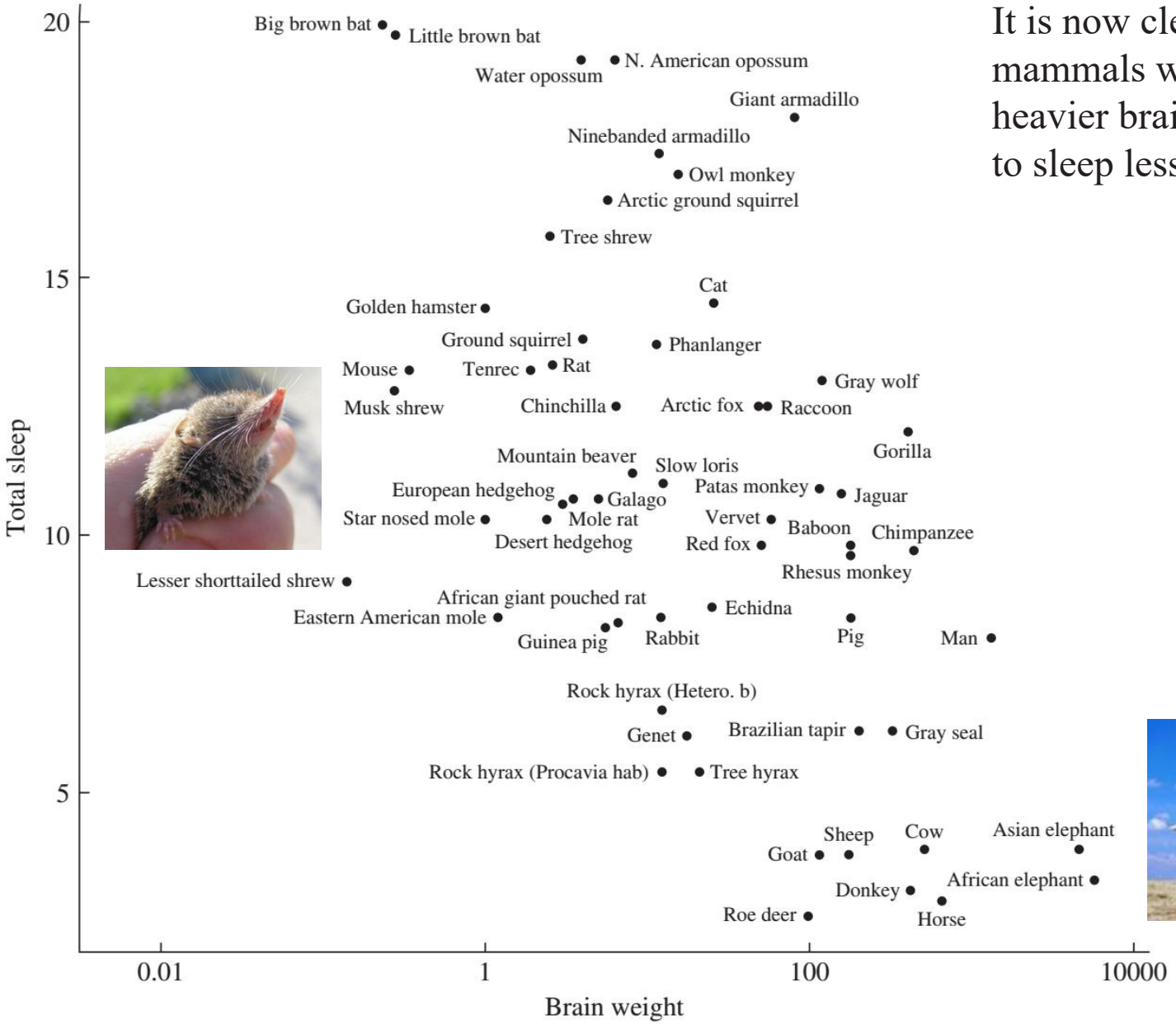
Two mammals with very large brains sleep very little, otherwise no relationships are apparent.

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!

Correlation - why do we try it?

When we make a set of measurements, it is instinct to try to correlate the observations with other results. We might wish

(1) to check that other observers' measurements are reasonable,

(2) to check that our measurements are reasonable,

(3) to test a hypothesis, perhaps one for which the observations were explicitly made,

(4) in the absence of any hypothesis, any knowledge, or anything better to do with the data, to find if they are correlated with other results in the hope of discovering some New and Universal Truth.

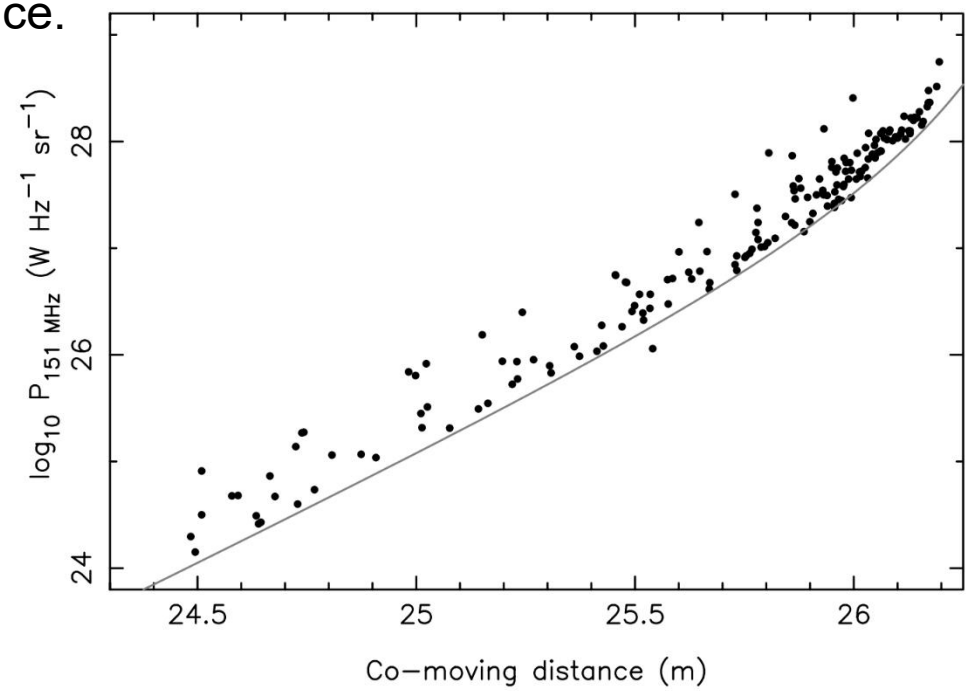
We are gonna do it – and we are going to fall into some deadly traps. We already have.

The fishing trip

Suppose that we have plotted something against something, on a Fishing Expedition.

1. Does the eye see much correlation? If not, formal testing for correlation is probably a waste of time. *The eyeball is an excellent statistical device.*

2. Could the apparent correlation be due to selection effects? Consider for instance the beautiful correlation obtained by Sandage (1972): 3CR radio luminosities vs distance.

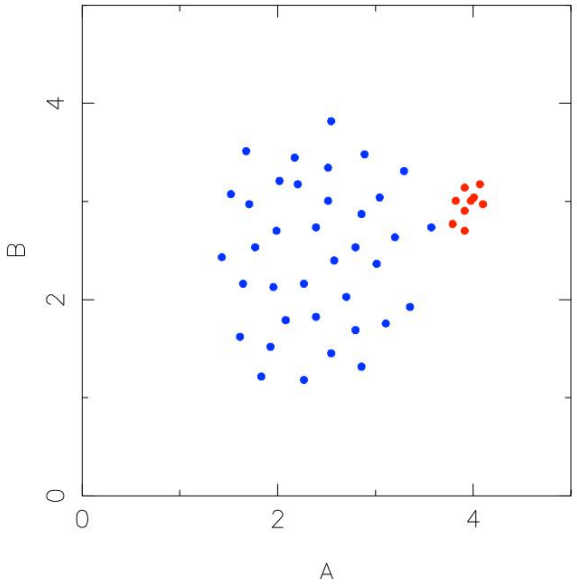


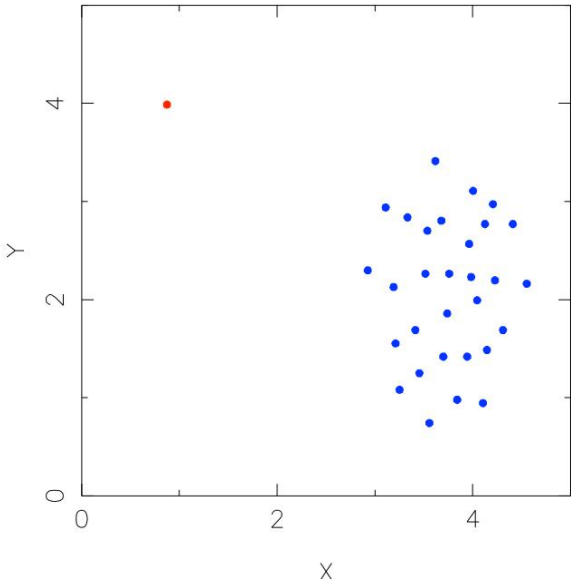
Radio luminosities of 3CR radio sources versus distance modulus

Still fishing …

3. If we are happy about (2), we can try formal calculation of the significance of the correlation. But, if there is a correlation, does the regression line (the fit) make sense?

4. If we are still happy - is the formal result realistic?
Rule of Thumb – *if* ***10% of the points*** *are grouped by themselves so that covering them with the thumb destroys the correlation to the eye, then we should doubt it.* Selection effects, data errors, or some other form of statistical conspiracy?





Suspect correlations: in each case formal calculation will indicate that a correlation exists to a high degree of significance!

Fishing, fishing …

5. If still confident, remember that
a correlation does not prove a causal connection. Examples:

- The price of fish in Billingsgate Market and the size of feet in China.
- Number of violent crimes in cities versus number of churches.
- The quality of student handwriting versus their height.
- Stock market prices and the sunspot cycle.
- In World War II, bombing accuracy was far greater when enemy fighter planes were present.
- Cigarette smoking versus lung cancer.
- Health versus alcohol intake…

1. Lurking third variables
2. Similar time scales
3. Causal connection…

There are ways of searching for intrinsic correlation between variables when they are known to depend mutually upon a third variable.

But… "known"???

Wilkinson & Pickett: *The Spirit Level*

`Correlations' show that higher income inequality correlates with higher crime rate, higher infant mortality, lower life expectancy, worse gender inequality, lower education standards, higher obesity rates……

Figure 5a: Wilkinson and Pickett's plot of inequality against homicide rates[33]

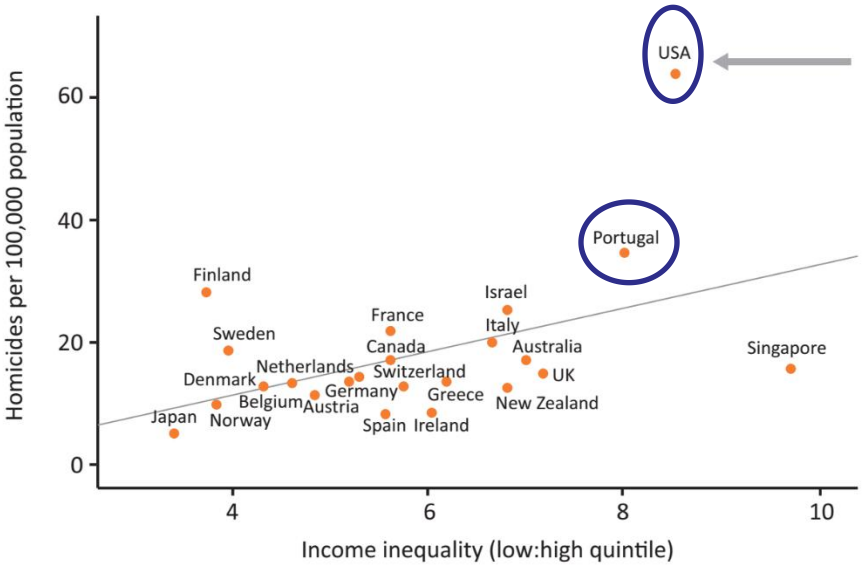
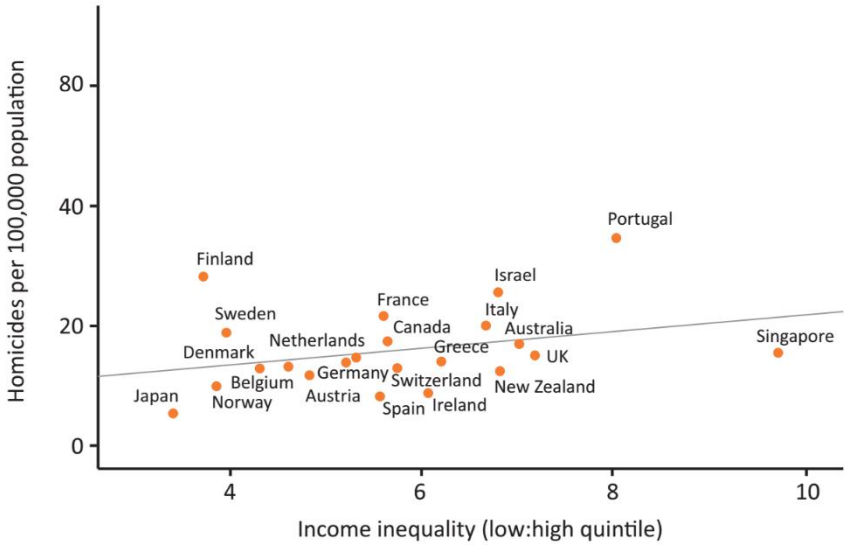


Figure 5b: Wilkinson and Pickett's plot of inequality against homicide rates, excluding the USA

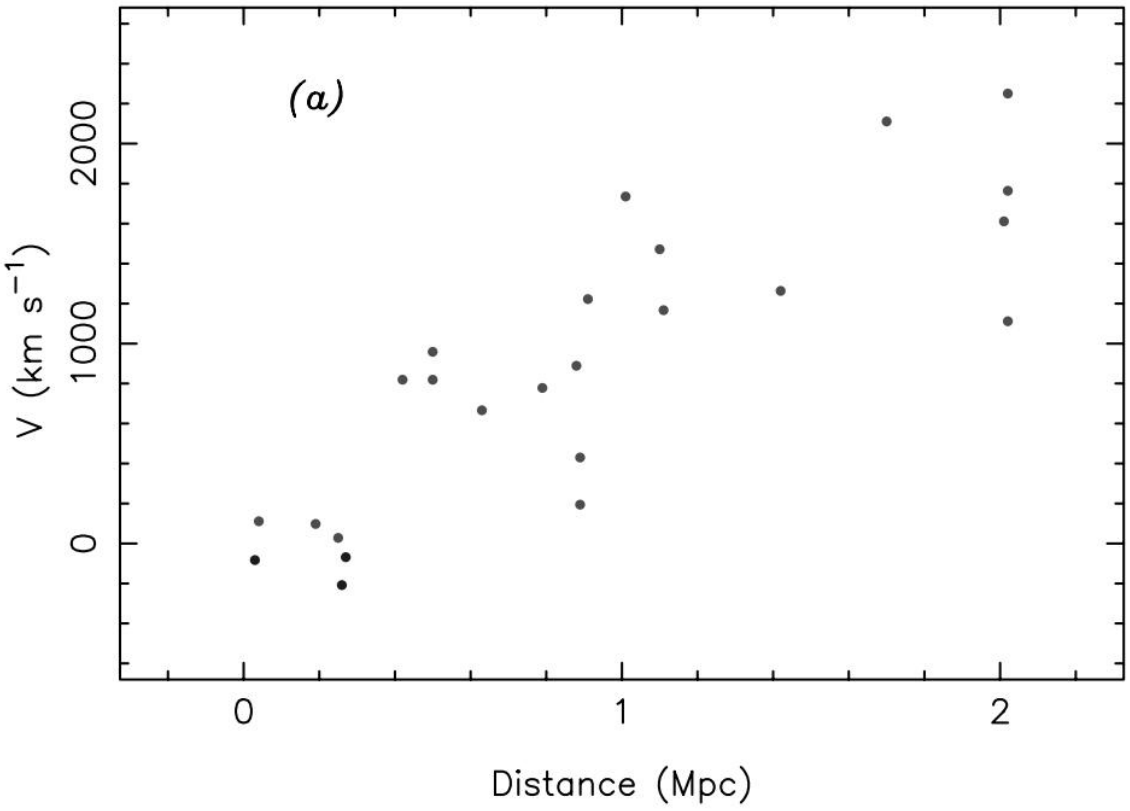


Critique by Peter Saunders: *Beware of False Prophets* shows that it is (statistical) garbage. The "correlations" are false or of no significance. The data are selective.

"Conclusion:There is no evidence of a significant association between the level of income inequality in a country and its homicide rate."

The end of the fishing trip – big fish are out there

Don’t get too discouraged by all the foregoing. Consider the example figure, a ragged correlation if ever there was one, although there are no nasty groupings of the type rejected by the Rule of Thumb.



An early Hubble diagram (Hubble 1936); recession velocities of a sample of 24 galaxies versus distance measure.

Formal correlation analysis later…

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.

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).

Comparing two samples

- continuing on hypothesis testing

Comparing two samples

Samples from distributions: Are they different? If so, how do they differ?

Samples are often drawn under different conditions, need inferences about their possible effects.

Primarily interest: those increase or decrease the average level of response.

Example. Cloud seeding:
Does it really increase precipitation?

- Some storms are seeded, others are not
- But precipitation varies widely from storm to storm
- A skeptic may not be convinced that the difference is due to anything but chance
- Develop statistical methods based on a stochastic model that treats the amounts of precipitation as random variables
- A process of randomization allows us to make inferences about treatment effects even if the observations are not modeled as samples from probability laws

Comparing two independent samples

Medical studies: a sample of subjects are assigned to a particular treatment, another independent sample assigned to a control (placebo安慰剂) treatment.

- randomly assigning individuals to the placebo and treatment groups

Control group: modeled as independent random variable with distribution F
Treatment group: independent of each other and of the controls with distribution G.

Let's focus on difference of location parameters (e.g. mean)

Methods based on Gaussian: *t* test

Treatment: X, μ_X, σ^2, Control: Y, μ_Y, σ^2, independent

MLE of $\mu_X - \mu_Y$ is

$$\overline{X} - \overline{Y} \sim N\left[\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right]$$

If σ^2 were known, a confidence interval for $\mu_X - \mu_Y$ could be based on

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Methods based on Gaussian: *t* test

Treatment: X, μ_X, σ^2, Control: Y, μ_Y, σ^2,

Confidence interval has the form

$$(\overline{X} - \overline{Y}) \pm z(\alpha/2)\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$$

But σ^2 needs to be estimated as **pooled sample variance**

$$s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2} \qquad s_X^2 = (n-1)\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Suppose that $X_1, \ldots, X_n$ are independent and normally distributed random variables with mean μ_X and variance σ^2, and that $Y_1, \ldots, Y_m$ are independent and normally distributed random variables with mean μ_Y and variance σ^2, and that the Y_i are independent of the X_i. The statistic

$$s_{\overline{X}-\overline{Y}} = s_p\sqrt{\frac{1}{n} + \frac{1}{m}}$$

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$$\frac{\Gamma\left[\frac{\nu+1}{2}\right]}{\sqrt{\pi\nu}\Gamma\left[\frac{\nu}{2}\right]}\left(1 + \frac{t'^2}{\nu}\right)^{-(\nu+1)/2}$$

$$\nu = n + m - 2$$

follows a t distribution with $m + n - 2$ degrees of freedom.

Methods based on Gaussian: *t* test

Example.

Two methods, A and B, were used in a determination of the latent heat of fusion of ice (Natrella 1963). The investigators wanted to find out by how much the methods differed. The following table gives the change in total heat from ice at $-.72°$C to water 0°C in calories per gram of mass:

Method A	Method B
79.98	80.02
80.04	79.94
80.02	79.98
80.04	79.97
80.03	79.97
80.03	80.03
80.04	79.95
79.97	79.97
80.05	
80.03	
80.02	
80.00	
80.02	

$$\overline{X}_A = 80.02 \qquad S_a = .024$$
$$\overline{X}_B = 79.98 \qquad S_b = .031$$

$$s_p^2 = \frac{12 \times S_a^2 + 7 \times S_b^2}{19} = .0007178$$
$$s_p = .027$$

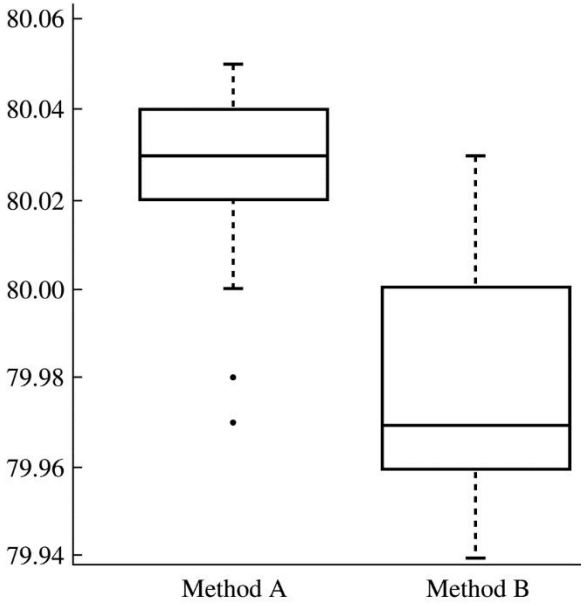
$$\overline{X}_A - \overline{X}_B = .04$$

$$s_{\overline{X}_A - \overline{X}_B} = s_p \sqrt{\frac{1}{13} + \frac{1}{8}} = .012$$

$$t_{19}(.025) = 2.093$$

95% confidence interval

$$(\overline{X}_A - \overline{X}_B) \pm t_{19}(.025) s_{\overline{X}_A - \overline{X}_B}, \text{ or } (.015, .065).$$



Methods based on Gaussian: *t* test

Hypothesis testing for two-sample problems. Null hypothesis: no treatment effect.

$$H_0: \mu_X = \mu_Y$$

Three common alternative hypotheses:

$$H_1: \mu_X \neq \mu_Y$$

Two-sided alternative
(more common in practice)

$$H_2: \mu_X > \mu_Y$$
$$H_3: \mu_X < \mu_Y$$

One-sided alternative

Test statistic

$$t = \frac{\overline{X} - \overline{Y}}{s_{\overline{X}-\overline{Y}}} \sim t_{m+n-2} \text{ (theorem)}$$

Same role in 2-sample comparison as is played by χ^2 in testing goodness of fit.

- Reject extreme *t* values, as rejecting large χ^2 values
- Knowing null distribution is *t* allows for a rejection region for a test at level α, as is χ^2 allowed for obtaining a rejection region for testing goodness of fit

Rejection regions:

$$\text{For } H_1, |t| > t_{n+m-2}(\alpha/2)$$
$$\text{For } H_2, t > t_{n+m-2}(\alpha)$$
$$\text{For } H_3, t < -t_{n+m-2}(\alpha)$$

Methods based on Gaussian: *t* test

Example. Latent heat of fusion of ice again.
To test H_0: $\mu_A = \mu_B$ versus a two-sided alternative,

$$t = \frac{\overline{X}_A - \overline{X}_B}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}} = 3.33$$

$t_{19}(.005) = 2.861$. The two-side test rejects at the level α=0.01.

(1) If two populations have different variance, estimate $\mathrm{Var}(\overline{X}-\overline{Y})$: $\frac{s_X^2}{n}+\frac{s_Y^2}{m}$
Using this as denominator, no longer *t* distribution.
But can be approximated by *t* distribution with d.o.f. as (round to nearest integer)

$$\mathrm{df} = \frac{[(s_X^2/n)+(s_Y^2/m)]^2}{\frac{(s_X^2/n)^2}{n-1}+\frac{(s_Y^2/m)^2}{m-1}}$$

(2) If underlying distributions are not normal and sample sizes are large, the use of the *t* distribution or the normal distribution is justified by CLT, probability levels of confidence intervals and hypothesis tests are approximately valid.

Methods based on Gaussian: *F* test

By analogous calculations, we can arrive at the *F* test for variances.

Again, Gaussian distributions are assumed.

The null hypothesis is H_0: $\sigma_x = \sigma_y$,

The data are X_i ($i = 1, \ldots, n$) and Y_i ($i = 1, \ldots, m$)

The test statistic is

$$\mathcal{F} = \frac{\sum_i (X_i - \overline{X})/(n-1)}{\sum_i (Y_i - \overline{Y})/(m-1)}.$$

following an *F* distribution with $n-1$ and $m-1$ degrees of freedom.

The testing procedure is the same as for *t* test.

Particular sensitive to the Gaussian assumption.

Nonparametric methods: the Mann-Whitney test

Nonparametric methods do not assume that the data follow any particular distributional form, often replacing the data by ranks. Why?

- Results are invariant under any monotonic transformation (in comparison, *p*-value of a *t* test may change on log scales).
- Using ranks has the effect of moderating the influence of outliers.

Mann-Whitney test (a.k.a. **Wilcoxon rank sum test**)

- We have $m + n$ experimental units to assign to a treatment and a control group.
- The assignment is made at random: Say, n units are randomly chosen and assigned to the control, and the remaining m units are assigned to the treatment.
- Null hypothesis: the treatment has no effect.
- If it is true, then any difference in the two outcomes is due to randomization

Nonparametric methods: the Mann-Whitney test

Test statistic: (the argument holds in the presence of ties)

- First, we group all $m + n$ observations together and rank them in order of increasing size.
- Calculate the sum of the ranks of those observations from the control group.
- If this sum is too small or too large, we will reject the null hypothesis.

A heuristic example. 4 subjects, 2 are **randomly** assigned to a treatment, other 2 to the control. Observed responses:

Treatment	Control
1 (1)	6 (4)
3 (2)	4 (3)

- Sum of ranks: control R=7, treatment =3, differ by chance?
- Calculate probability of such a discrepancy if treatment has no effect at all, difference entirely due to particular randomization — the null hypothesis

Nonparametric methods: the Mann-Whitney test

Key idea:

- we can explicitly calculate the distribution of R under the null hypothesis: every assignment of ranks to observations is equally likely, enumerate all 4! = 24 such assignments.
- In particular, each of the $\binom{4}{2} = 6$ assignments of ranks to the control group is equally likely:

Ranks	R
{1, 2}	3
{1, 3}	4
{1, 4}	5
{2, 3}	5
{2, 4}	6
{3, 4}	7

- Under null hypothesis, R's (null) distribution is

r	3	4	5	6	7
$P(R = r)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$

$$P(R = 7) = \frac{1}{6}$$

This discrepancy would occur one time out of six purely on the basis of chance.

Nonparametric methods: the Mann-Whitney test

More practically: Say, there are n observations in treatment group, m in control.

- If null hypothesis holds, every assignment of ranks to the $m + n$ observations is equally likely, each of the $\binom{m+n}{m}$ possible assignments of ranks to the control group is equally likely.

Nonparametric methods: the Mann-Whitney test

More practically: Say, there are n observations in treatment group, m in control.

- If null hypothesis holds, every assignment of ranks to the $m + n$ observations is equally likely, each of the $\binom{m+n}{m}$ possible assignments of ranks to the control group is equally likely.
- For each assignment, we calculate the sum of the ranks and thus determine the null distribution of the test statistic -- the sum of the ranks of the control group
    - No assumption that data from control/treatment are samples from a probability distribution. Probability kicks in due to random assignment

Nonparametric methods: the Mann-Whitney test

More practically: Say, there are n observations in treatment group, m in control.

- If null hypothesis holds, every assignment of ranks to the $m + n$ observations is equally likely, each of the $\binom{m+n}{m}$ possible assignments of ranks to the control group is equally likely.
- For each assignment, we calculate the sum of the ranks and thus determine the null distribution of the test statistic -- the sum of the ranks of the control group
  - No assumption that data from control/treatment are samples from a probability distribution. Probability kicks in due to random assignment
  - Rank sum is easy to compute and sensitive to a treatment. But any other test statistic can be used and its null distribution computed in the same fashion. Also, **its null distribution has to be computed only once and tabled**.
  - The sum of the two rank sums is $1+2+\ldots+(m+n)=[(m+n)(m+n+1)/2]$, knowing one rank sum tells us the other.

Nonparametric methods: the Mann-Whitney test

More practically: Say, there are n observations in treatment group, m in control.

- If null hypothesis holds, every assignment of ranks to the $m + n$ observations is equally likely, each of the $\binom{m+n}{m}$ possible assignments of ranks to the control group is equally likely.
- For each assignment, we calculate the sum of the ranks and thus determine the null distribution of the test statistic -- the sum of the ranks of the control group
  - No assumption that data from control/treatment are samples from a probability distribution. Probability kicks in due to random assignment
  - Rank sum is easy to compute and sensitive to a treatment. But any other test statistic can be used and its null distribution computed in the same fashion. Also, **its null distribution has to be computed only once and tabled**.
  - The sum of the two rank sums is $1+2+\ldots+(m+n)=[(m+n)(m+n+1)/2]$, knowing one rank sum tells us the other.
- Tables in terms of rank sum of the smaller group, or smaller of two rank sums
- Let n_1=smaller sample size, R=its rank sum, $R'=n_1(m+n+1)-R$, then critical value $R^* = \min(R, R')$ is given in tables.

Nonparametric methods: the Mann-Whitney test

If there are only a small # of ties, tied observations are assigned average rank; then significance levels are not greatly affected.

Example. Latent heats of fusion of ice. Sample sizes=13 and 8, small, no prior knowledge validating Gaussian assumption, safer to use nonparametric methods.

Method A	Method B
79.98	80.02
80.04	79.94
80.02	79.98
80.04	79.97 ←
80.03	79.97 ←
80.03	80.03
80.04	79.95
79.97 ←	79.97 ←
80.05	
80.03	
80.02	
80.00	
80.02	

Method A	Method B
7.5	11.5
19.0	1.0
11.5	7.5
19.0	4.5 ←
15.5	4.5 ←
15.5	15.5
19.0	2.0
4.5 ←	4.5 ←
21.0	
15.5	
11.5	
9.0	
11.5	

不依赖于正态假设，对离群值不敏感，即使正态性假设成立，其势也接近*t*检验，因而广泛使用，尤其小样本情况下

The sum of the ranks of the smaller sample is $R = 51$. $R' = 8(8+13+1) - R = 125$, $R^* = 51$.

Table: **53** is the critical value for a two-tailed test with $\alpha = .01$.
Reject null hypothesis at this significance level.

表 8　威尔科克森和曼恩－惠特尼检验中较小秩和的临界值

n_2	双边检验的 α	单边检验的 α	n_1(较小的样本)																		
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	0.20	0.10		3	7																	
	0.10	0.05			6																	
	0.05	0.025																				
	0.01	0.005																				
4	0.20	0.10		3	7	13																
	0.10	0.05			6	11																
	0.05	0.025				10																
	0.01	0.005																				
5	0.20	0.10		4	8	14	20															
	0.10	0.05		3	7	12	19															
	0.05	0.25			6	11	17															
	0.01	0.005					15															
6	0.20	0.10		4	9	15	22	30														
	0.10	0.05		3	8	13	20	28														
	0.05	0.025			7	12	18	26														
	0.01	0.005				10	16	23														
7	0.20	0.10		4	10	16	23	32	41													
	0.10	0.05		3	8	14	21	29	39													
	0.05	0.025			7	13	20	27	36													
	0.01	0.005				10	16	24	32													
8	0.20	0.10		5	11	17	25	34	44	55												
	0.10	0.05		4	9	15	23	31	41	51												
	0.05	0.025		3	8	14	21	29	38	49												
	0.01	0.005				11	17	25	34	43												
9	0.20	0.10	1	5	11	19	27	36	46	58	70											
	0.10	0.05		4	*10	16	24	33	43	54	66											
	0.05	0.025		3	8	14	22	31	40	51	62											
	0.01	0.005			6	11	18	26	35	45	56											

(续)

n_2	双边检验的 α	单边检验的 α	n_1(较小的样本)																		
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
10	0.20	0.10	1	6	12	20	28	38	49	60	73	87										
	0.10	0.05		4	10	17	26	35	45	56	69	82										
	0.05	0.025		3	9	15	23	32	42	53	65	78										
	0.01	0.005			6	12	19	27	37	47	58	71										
11	0.20	0.10	1	6	13	21	30	40	51	63	76	91	106									
	0.10	0.05		4	11	18	27	37	47	59	72	86	100									
	0.05	0.025		3	9	16	24	34	44	55	68	81	96									
	0.01	0.005			6	12	20	28	38	49	61	73	87									
12	0.20	0.10	1	7	14	22	32	42	54	66	80	94	110	127								
	0.10	0.05		5	11	19	28	38	49	62	75	89	104	120								
	0.05	0.25		4	10	17	26	35	46	58	71	84	99	115								
	0.01	0.005			7	13	21	30	40	51	63	76	90	105								
13	0.20	0.10	1	7	15	23	33	44	56	69	83	98	114	131	149							
	0.10	0.05		5	12	20	30	40	52	64	78	92	108	125	142							
	0.05	0.025		4	10	18	27	37	48	60	73	88	103	119	136							
	0.01	0.005			7	*13	22	31	41	53	65	79	93	109	125							
14	0.20	0.10	1	*8	16	25	35	46	59	72	86	102	118	136	154	174						
	0.10	0.05		*6	13	21	31	42	54	67	81	96	112	129	147	166						
	0.05	0.025		4	11	19	28	38	50	62	76	91	106	123	141	160						
	0.01	0.005			7	14	22	32	43	54	67	81	96	112	129	147						
15	0.20	0.10	1	8	16	26	37	48	61	75	90	106	123	141	159	179	200					
	0.10	0.05		6	13	22	33	44	56	69	84	99	116	133	152	170	192					
	0.05	0.025		4	11	20	29	40	52	65	79	94	110	127	145	164	184					
	0.01	0.005			8	15	23	33	44	56	69	84	99	115	133	151	171					
16	0.20	0.10	1	8	17	27	38	50	64	78	93	109	127	145	165	185	206	229				
	0.10	0.05		6	14	24	34	46	58	72	87	103	120	138	156	176	197	219				
	0.05	0.025		4	12	21	30	42	54	67	82	97	113	131	150	169	190	211				
	0.01	0.005			8	15	24	34	46	58	72	86	102	119	136	155	175	196				

Linear least squares

Linear least squares

The most common (but by no means only) method for determining the parameters in curve-fitting problems.

A straight line is fit to (y_i, x_i), where $i = 1, \ldots, n$;

- y is called **dependent/response variable**, x is **independent/predictor variable**, predict y from x. We minimize

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Procedure is not symmetric in y and x!

To find β_0 and β_1, we calculate

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

Linear least squares vs. correlation

Setting the partial derivatives to 0, the minimizers satisfy

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \qquad \sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Let's introduce:

$$s_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad s_{yy} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Correlation between x's and y's is $r = \dfrac{s_{xy}}{\sqrt{s_{xx} s_{yy}}}$

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} \qquad r = \hat{\beta}_1 \sqrt{\frac{s_{xx}}{s_{yy}}}$$

Correlation is zero if and only if the slope is zero.

The concept of regression

After some manipulation and standardizing the variables,

$$\frac{\hat{y} - \bar{y}}{\sqrt{s_{yy}}} = r \frac{x - \bar{x}}{\sqrt{s_{xx}}}$$

Interpretation:

- Suppose $r > 0$ and predictor variable x is one standard deviation > x's average
- Then predicted value of y is r standard deviations > y's average
- $r \leq 1$, in units of standard deviations, y is closer to its average than x - **Regression**

*Regression Analysis*是1886年英国遗传学家和统计学家Sir Francis Galton提出

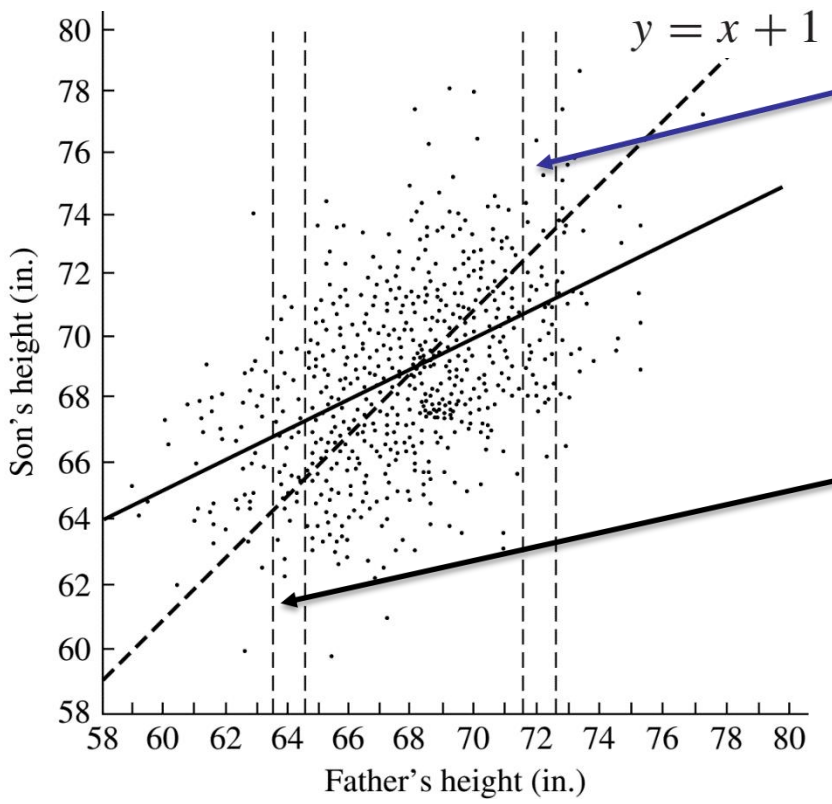
- 他研究种子及其后代的尺寸、父亲及其儿子的身高之间的关系
- 在《身高遗传中的平庸回归》论文中首先使用术语“回归”
- 子女身高与父母有关，高大的父母子女一般也较高大，反之亦然，但此趋势不会向两个极端无限发展
- Galton对1074对父母及其一个成年儿子的身高分析后发现，如果父辈身材高大，后代往往比父辈矮小一些，父辈矮小，后代比父辈高大一些

“**regression towards mediocrity**”

The concept of regression

(Freedman, Pisani, Purves, 1998) The heights of 1078 pairs of fathers and sons. The fathers' average height is 67.7 in. with a standard deviation of 2.74 in.; the sons' average and standard deviation are 68.7 in. and 2.81 in., respectively; the correlation coefficient is 0.501.

Notice how the prediction son's height = father's height + 1 under-predicts on the left and over-predicts on the right.



Fathers' heights are 72 in, sons' average is 71 in. Regression line

$$\frac{\hat{y} - 68.7}{2.81} = .5 \times \frac{x - 67.7}{2.74}$$

predicts sons' height = 70.9 in.

Fathers' heights are 64 in., sons' average is 67 in. Regression line predicts 66.8 in.

The concept of regression

Example (*Tooooooooold!!!*).
Two exams are given in a course. The scores of a student on the mid-term and final exams, X and Y, are jointly distributed.

Suppose that the exams are scaled to have the same means $\mu = \mu_X = \mu_Y$ and standard deviations $\sigma = \sigma_X = \sigma_Y$. Then, the correlation $\rho = \sigma_{XY}/\sigma^2$ and the best linear predictor $\hat{Y} = \mu + \rho(X - \mu)$.

By the equation $$\hat{Y} - \mu = \rho(X - \mu)$$
we predict that student's score on the final exam to differ from the overall mean μ by less than did the score on the mid-term.

In case of positive correlation:

-- Encouraging for students below average, bad news for those above average

This phenomenon is often referred to as regression to the mean.

From Lec 5, Predictions

Statistical properties of estimated slope and intercept

Reliability of the slope and intercept in the presence of "noise"?

Need a statistical model. The simplest one (standard statistical model)

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, \ldots, n$$

Here the e_i are independent random variables with $E(e_i) = 0$ and $\text{Var}(e_i) = \sigma^2$.

Important conclusion A:

Under the assumptions of the standard statistical model, the least squares estimates are unbiased: $E(\hat{\beta}_j) = \beta_j$, for $j = 0, 1$.

Important conclusion B:
We define the **residual sum of squares** (**RSS**, 残差平方和)

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

then an unbiased estimate of σ^2 is

$$s^2 = \frac{\text{RSS}}{n-2}$$

五种一元线性回归

a) 常规线性回归OLS($Y|X$)，最小化Y方向回归线线距离

b) 逆回归OLS($X|Y$)，最小化X方向距离

c) 正交回归线orthogonal regression (OR)，最小化垂直距离

d) 简化主轴回归reduced major-axis regression (RMA)，X,Y两方向距离

e) 平分线bisector，平分OLS($Y|X$) 和OLS($X|Y$)

- 如果五种方法差异不大，可以使用OLS($Y|X$)
- 如果一个明显是自变量，一个明显是因变量，应使用OLS($Y|X$)
- 如果从一个变量预测另一个变量，也应使用OLS($Y|X$)
- 如果研究目的是了解变量间的基本关系，处理对称变量的三种方法OR，RMA，平分线都可用，但普遍认为平分线法值得推荐

- Isobe, Feigelson et al. Linear regression in astronomy, 1990, ApJ, 364, 104
- Feigelson & Babu, Linear regression in astronomy II, 1992, ApJ, 397, 55

Suggestions for your further study

更多统计方法及其在天文学中更多的实际应用（宜常备的案头参考书）：

Practical statistics for astronomers, Wall & Jenkins

- 优点：内容极广、材料很新，多有与天文学的结合；对各种统计方法和思想有高屋建瓴式的评论（优缺点，适用场合，天文学家的视角）；书末对有用的统计和天文资料多有评述。

- 缺点：1. 貌似浅显（公式少），其实是写给已经非常了解统计学的人，字里行间信息量太大，不能作为第一本书；2. 英语表达往往过于俚语化，为求生动而多用典故，不熟悉英美国家语言和文化背景者读得似懂非懂。

- 我们的课程内容和John A. Rice的书提供了很好的基础；可以在此基础上利用该书追踪自己所需的更为深入具体的资料 -- **Appendix A极为有用**！

Data reduction and error analysis for the physical sciences, Bevington & Robinson

- 优点：简洁清晰，实用至上，初学者较易入手

- 缺点：过多着力于数据拟合技术的讨论，内容较窄，很多深入和高等内容缺失，内容略显陈旧

- 主教材：*Data Reduction and Error Analysis for the Physical Sciences*
  - by PR Bevington & DK Robinson, 2003, 3rd edn, 制本厂有影印本
  - 内容精炼，简洁清晰，以实用至上的风格广受赞誉
- 参考书一：*Mathematical Statistics and Data Analysis*
  - by John A. Rice, 2007, 3rd edn （有电子版）
  - 系统的教科书，内容丰富而深入，实例众多，宜常备的基础统计书
- 参考书二：*Practical Statistics for Astronomers*
  - by JV Wall and CR Jenkins, 2012, 2nd edn
  - 实用、全面、易读、跟进时代，宜常备的天文统计书
- *Numerical Recipes in C/C++/FORTRAN*, by WH Press et al.
  - 最经典的天文工具箱，在学术界有圣经般的地位

Modern Cosmology, by S Dodelson

- 末章集中讲解宇宙学中常用的统计方法，宇宙学数据分析入门必读

Bayesian Logical data Analysis for the Physical Sciences

- 如果你需要进一步学习贝叶斯理论，可以以此为起点

Rao的结束语

一切的知识，归根结底都是历史。

一切的科学，抽象看来都是数学。

一切的判断，寻根问底都是统计学。