Historia Rachunku Prawdopodobieństwa i Statystyki
WYKŁAD V: Quetelet. Legendre i Gauss. CTG Laplace'a

MiNI PW

Adolf Quetelet (1796-1874)

Adolf Qutelet: dobrze wykształcony matematyk, pierwszy doktorat z nauk ścisłych Uniwersytetu w Gandawie.
Członek Królewskiej Akademii Nauk i Literatury od 1820 roku.
Założyciel Królewskiego Obserwatorium w Brukseli (1828).
Bibliografia jego prac zawiera ponad 300 pozycji.
Głowne dzieło *O człowieku i rozwoju jego cech. Rozprawa o fizyce społecznej* (dwa tomy opublikowane w 1835 roku). Drugie wydanie *Rozprawa o fizyce społecznej* (1869)
Fizyka społeczna: analiza ilościowa danych dotyczących społeczeństw.

Adolf Quetelet (1796-1874), statystyk, socjolog, astronom

Adolf Quetelet (1796-1874)

Głowne osiągnięcia:

- Próba wprowadzenia metody ilorazowej Laplace'a i dyskusja metodologiczna z tym związana;
- koncepcja człowieka przeciętnego (*l'homme moyen*)
- wkład do analizy danych (ilościowa analiza danych społecznych)
- dopasowanie rozkładu normalnego do danych
- prawo przyczyn przypadkowych

Metoda ilorazowa Laplace'a

Laplace: metoda oceny wielkości populacji na podstawie przeskalowania danych z jednego departamentu.

$$\frac{N}{N_D} = \frac{U}{U_D},$$

gdzie:
N, N_D - liczby mieszkańców kraju i ustalonego departamentu odpowiednio;
U, U_D - liczby urodzin w kraju i departamencie odpowiednio.
Zatem

$$N = N_D \times \frac{U}{U_D},$$

czyli znając liczbę urodzin w departamencie i kraju oraz liczbę mieszkańców w departamencie byliśmy w stanie oszacować liczbę N.
Podobnie

$$N = N_D \times \frac{Z}{Z_D},$$

gdzie Z i Z_D odpowiednie liczby zgonów.

Krytyka de Kevenberga: Zmienność liczby zgonów i urodzin w departamentach jest bardzo duża i zależy od wielu czynników (gęstość zaludnienia, wysokość nad poziomem morza..) więc dane z jednego departamentu nie są *reprezentatywne*. Można wprawdzie zmodyfikować wzór

$$N = \frac{1}{K} \sum_{k=1}^{K} N_{D,k} \times \frac{Z}{Z_{D,k}},$$

ale wdg Kevenberga prowadziłoby to do wybrania tylu departamentów, że metoda straciłaby swoją użyteczność w porównaniu z pełnym spisem.
AQ zgodził sie z Kavenbergiem i odstąpił od propagowania metody ilorazowej.
Kwestia złych metodologii i błędnych badań ...

Koncepcja człowieka przeciętnego (*l'homme moyen*)

Przeciętność dla AQ oznaczała bycie blisko wartości średniej charakterystyki w rozpatrywanej populacji.
Koncepcja człowieka przeciętnego ewoluowała w czasie.

Koncepcja człowieka przeciętnego (*l'homme moyen*)

Koncepcja ilościowego spojrzenia na społeczeństwo i prawa społeczne rozwijana w tamtym czasie.
Jedna z nagród Królewskiego Towarzystwa w Brukseli za rok 1828: *Przedstawić teorię matematyczną ludzi i zwierząt traktowanych jako motory i maszyny*.
Zaproponował koncepcję człowieka przeciętnego jako posiadacza średnich wartości cech policzonych dla pewnej społeczności. Miało to dotyczyć nie tylko cech antropometrycznych ale również kategorii moralnych i obyczajowych.
If one individual at any given epoch of society possesed all the qualities of the average man, he would represent all that is great, good or beautiful

Krytyka

Koncepcja człowieka przeciętnego zyskała w tamtych czasach dużą popularność, podobnie jak teraz.
Jednak:
Recenzenci mieli wątpliwości co do rozciągnięcia idei 'normalności' na sferę moralną:
Jeśli chodzi o sferę moralności taka przeciętność byłaby nie do zaakceptowania
A. Cournot: Idea określania typowości poprzez liczenie średniej nie ma sensu, gdyż obiekt średni nie musi być elementem badanej klasy.
Argument: Trójkąt mający boki będące średnimi boków trójkątów prostokątnych nie musi być trójkątem prostokątnym.

Koncepcja normalności

Sam AQ widział, ze żeby porównywać wartość ze średnią, konieczna jest miara zmienności wokół średniej.
Wprowadził, na podstawie odstępstwa od średniej, wskaźnik względnej skłonności do popełnienia przestepstwa w zależności od wieku ...
Koncepcja normalności: jeśli odchylenie jakiejeś jednostki od średniej było nieistotne, AQ traktował ją jako 'normalną'...
Co jest normalne ?
American Psychiatric Association diagnostic and statistical manual
DSM-V: znaczne rozszerzenie liczby zaburzeń traktowanych jako choroby
..
Problem wielokrotnego testowania w przypadku wielu cech !!!

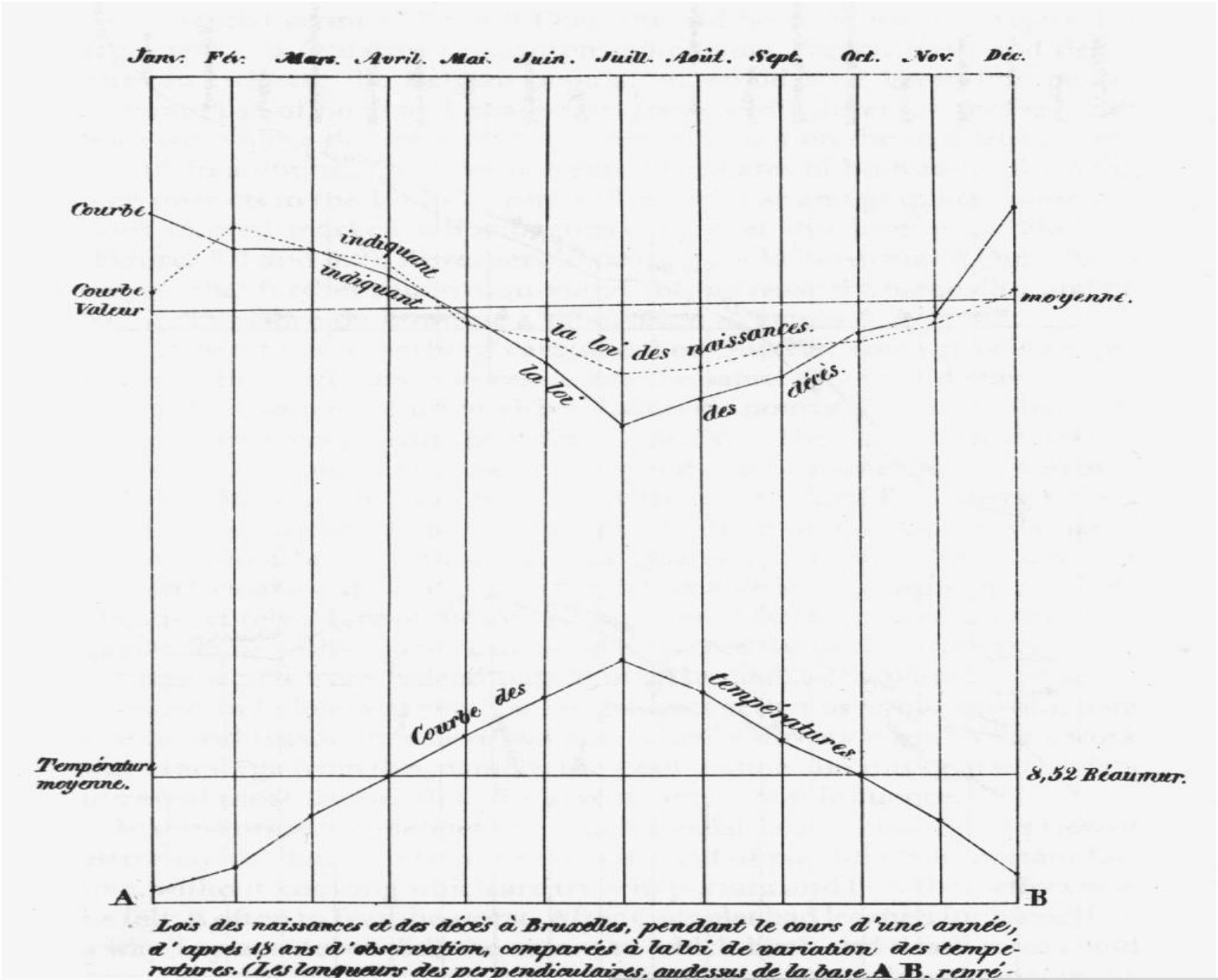
Wkład do analizy danych (ilościowa analiza danych społecznych)

AQ był zafascynowany metodami syntetycznego spojrzenia na dane i możliwościami wykresów.
Nie wyszedł ponad reprezentację dwóch zmiennych jednocześnie, ale pomijając to jego ciekawość była nieograniczona:
zgony i urodziny versus miesiąc, miasto, temperatura, pora dnia;
zgony versus wiek, zawód, miejsce zamieszkania;
zgony w więzieniach i szpitalach;
statystyki dotyczące pijaństwa, przestępczości i chorób psychicznych.
Zadziwiając regularności: liczba małżenstw
(kobieta $\geqslant 60$, mężczyzna $\leqslant 30$ w Belgii w latach 1841 - 1845):

$$7, 6, 8, 5, 5$$

odpowiednio.
To doprowadziło go do prawa przyczyn przypadkowych.

Reprezentacje graficzne: narodziny i zgony vs czas



Reprezentacje graficzne Minard

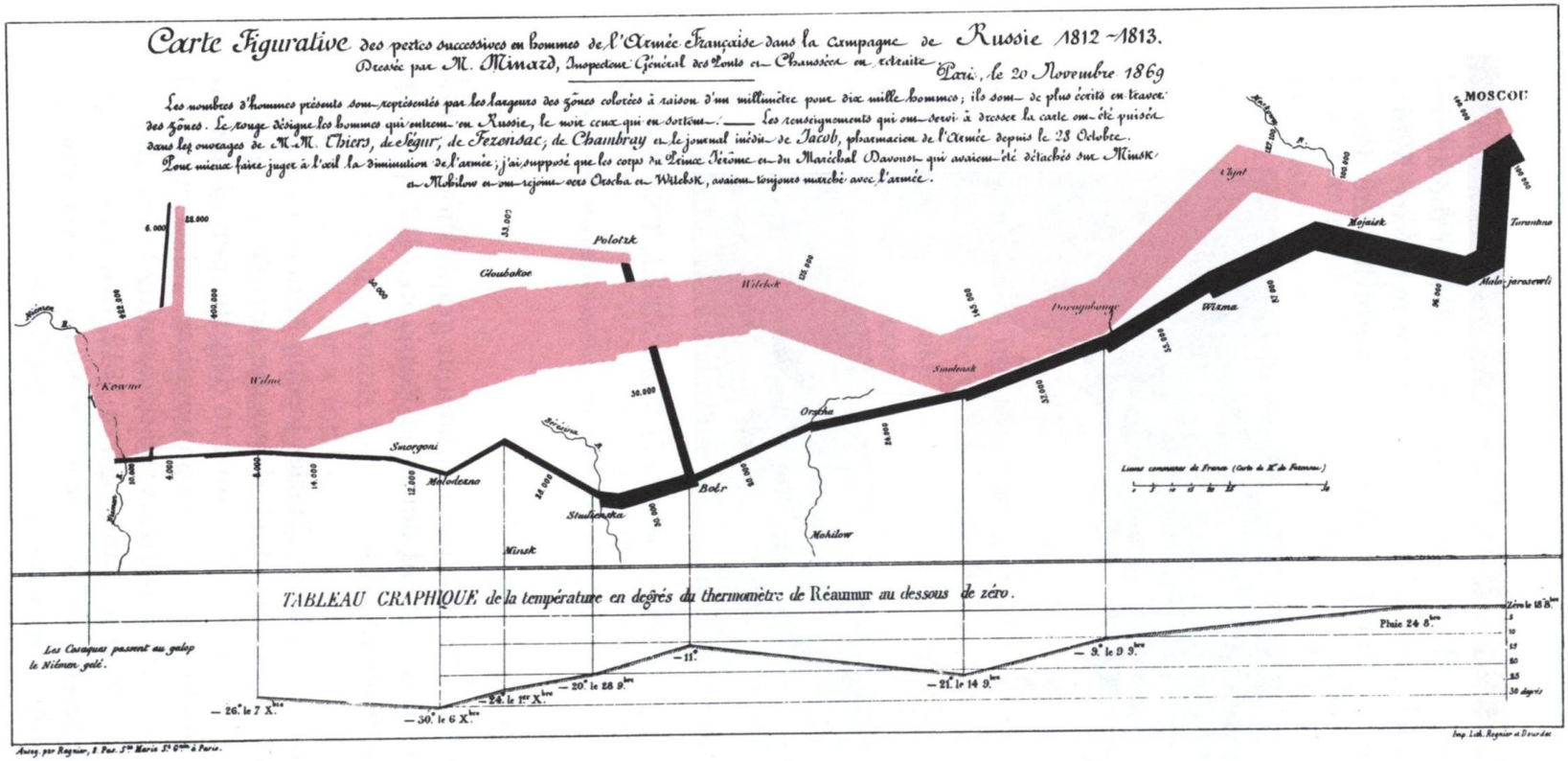


Figure 3.12 Napoleon's march to and retreat from Moscow. *(Source: Edward R. Tufte,* The Visual Display of Quantitative Information, *Cheshire, CT: Graphic Press, 1983.)*

Autor: Karol Minard

Berezyna, listopad 1812

Modelowanie zależności

Próba modelowania związku między wagą i wzrostem

$$y + \frac{y}{1000(T-y)} = ax + \frac{t+x}{1+(4/3)x}$$

t, T - wzrost w momencie urodzin i zgonu;
x- waga, y- wzrost, a - stała zależna od populacji.

Stabilność charakterystyk populacji

Teza AQ:
Jeśli związki przyczynowe nie ulegają zmianie to średnie charakterystyki w populacji (nawet niejednorodnej) się nie zmieniają.
Wprowadził tzw indeks Queteleta otyłości (od 1972 roku zwany BMI).

Prawo przyczyn przypadkowych

Ogólne prawo rządzące wszechświatem: prawo przyczyn przypadkowych (*loi des causes accidentalles*).
Zależności są wynikiem działania dwóch sił: sił stałych (*naturelles*) i sił zakłócających (*perturbatrices*). Siły stałe określają stan przeciętny (*etat moyen*), a siły zakłócające (losowe) odpowiedzialne są za odstępstwa. W przypadku wzrostu: stan przeciętny: wzrost średni, siły zakłocające: klimat, odżywianie itp.
W przypadku zjawisk społecznych siła zakłocającą może być wolna wola. Wpływ wolnej woli na przebieg zjawisk społecznych ogranicza się do roli przyczyny przypadkowej i jej wypadkowa maleje w przypadku dużych populacji.
Buckle *History of Civilisation in England* 1857
Quteletismus

Dopasowanie rozkładu normalnego do danych: Quetelet

Qutelet zrozumiał znaczenie CTG Laplace'a dla analizy danych społecznych. Przykład analizy: obwód klatki piersiowej 5738 zołnierzy szkockich (najczęsciej używane dane w XIX wieku) (moda 40 cali=102 cm).
AQ wykorzystał swoje, sprytnie policzone, tablice rozkładu normalnego oparte na losowaniu 999 kul z 'dużej' urny zawierającej jednakową liczbę kul białych i czarnych (de facto Bin(999,1/2)).
Metoda opierała się na policzeniu skumulowanych prawdopodobieństw normalnych

$$S(r) = P(500 \leqslant X < 500 + r)$$

policzeniu analogicznego skumulowanego prawdopodobieństwa $\hat{p}$ i

$$\hat{r} = S^{-1}(\hat{p})$$

Natępnie interpolacja liniowa $\hat{r} \longrightarrow \tilde{r}$ i wyliczenie odpowiednich prawdopodobieństw obwodu klatki na podstawie $S(\tilde{r})$.

Dane: obwody klatki piersiowej

MESURES de la POITRINE.	NOMBRE d'hommes.	NOMBRE PROPORTIONNEL.	PROBABILITÉ d'après L'OBSERVATION.	RANG dans LA TABLE.	RANG d'après le CALCUL.	PROBABILITÉ d'après LA TABLE.	NOMBRE D'OBSERVATIONS calculé.
Pouces.							
33	3	3	0,5000			0,5000	7
34	18	31	0,4993	52	50	0,4993	29
35	81	141	0,4964	42,5	42,5	0,4964	110
36	185	322	0,4823	33,5	34,5	0,4854	323
37	420	732	0,4501	26,0	26,5	0,4531	732
38	740	1305	0,3769	18,0	18,5	0,3799	1333
39	1073	1867	0,2464	10,5	10,5	0,2466	1858
			0,0597	2,5	2,5	0,0628	
40	1079	1882	0,1285	5,5	5,5	0,1359	1987
41	934	1628	0,2913	13	13,5	0,3034	1675
42	658	1148	0,4061	21	21,5	0,4130	1096
43	370	645	0,4706	30	29,5	0,4690	560
44	92	160	0,4866	35	37,5	0,4911	221
45	50	87	0,4953	41	45,5	0,4980	69
46	21	38	0,4991	49,5	53,5	0,4996	16
47	4	7	0,4998	56	61,8	0,4999	3
48	1	2	0,5000			0,5000	1
	5738	1,0000					1,0000

Figure 5.3. Quetelet's analysis fitting a normal distribution to data on the chest circumferences of Scottish soldiers. Column 1 gives the chest circumference in inches; columns 2 and 3 give the frequency and relative frequency distributions for 5738 individuals, columns 4–7 give the details of Quetelet's calculations (see text), and column 8 gives the fitted relative frequency distribution. (From Quetelet, 1846, p. 400.)

Henry Buckle 'Historia cywilizacji w Anglii'

Henry Buckle 'Historia cywilizacji w Anglii'

Henry Buckle (1821-1862) : historia rządzi sie materialistycznymi prawami, wydarzenia nastepują w konsekwencji wydarzeń poprzednich, a nie w w wyniku arbitralnych decyzji jednostek czy działania sił nadprzyrodzonych. Klimat i dostępność żywności są podstawymi czynnikami wpływającymi na rozwój cywilizacji.
Istnieje 'konieczność dziejowa', ale nie należy jej przyśpieszać (inaczej sądził Marks)
Koncepcja społeczeństwa jaka zbioru dużej liczby jednostek rządzącego się pewnymi ogólnymi prawami propagowana przez Queteleta i Buckle'a została wykorzystana przez fizyków do stworzenia podstaw **fizyki statystycznej**.

Główne postaci

- Rudolph Claussius (1822-1866)
- James Maxwell (1831-1879)
- Ludwig Boltzmann (1844 -1906)
- Marian Smoluchowski
- Josiah Gibbs (1839-1903)

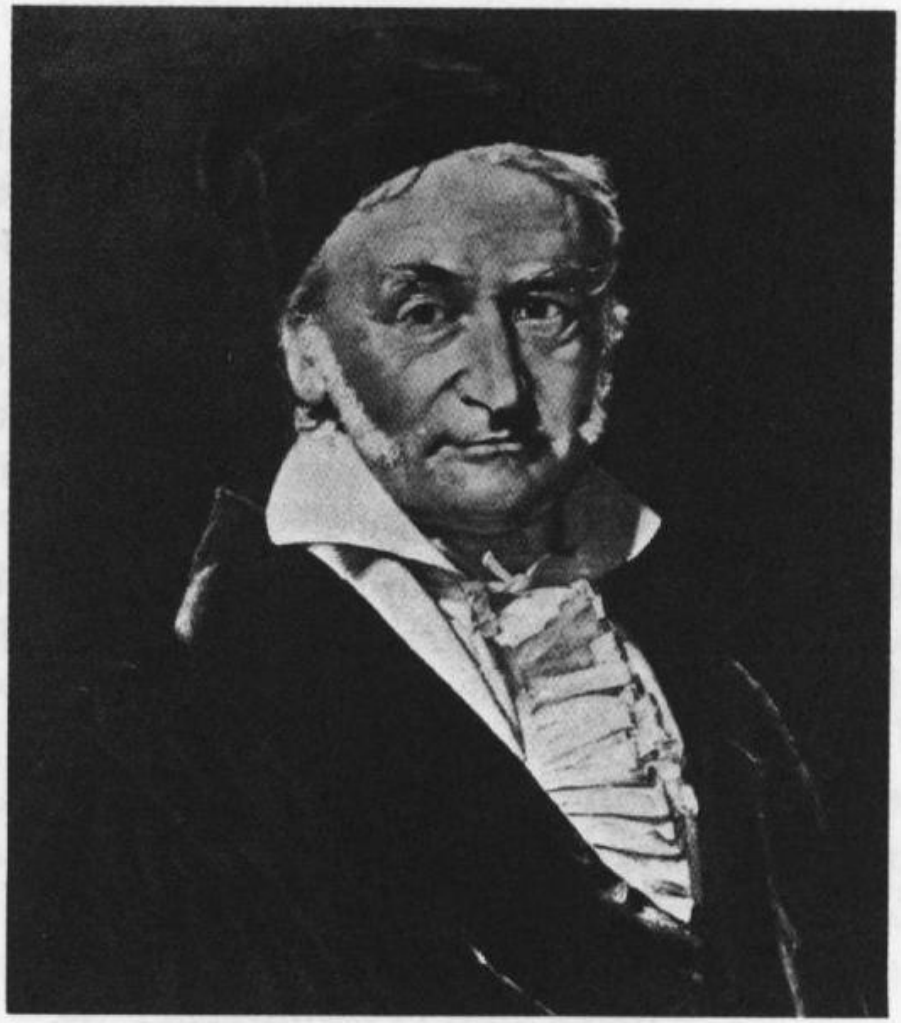
Legendre i metoda najmniejszych kwadratów

Adrien Marie Legendre (1752-1833)
Pierwsza wzmianka o MNK w pracy o wyliczaniu orbit komet.
Później wykorzystana przez niego przy liczeniu długości południka (od równika do bieguna północnego, *meridian quadrant*) przechodzącego przez Paryż (1805).
Na niej był oparty wzorzec metra (metr$= 10^{-7}$ długości). Dopasowanie równania parametrycznego do danych empirycznych: szukamy minimum funkcji kryterialnej.
MNK traktowana jako metoda analizodanowa, bez modelu probabilistycznego.
W szczególności zdawał sobie sprawę, że reguła brania średniej obserwacji to przypadek szczególny metody MNK.

Legendre i metoda najmniejszych kwadratów cd.

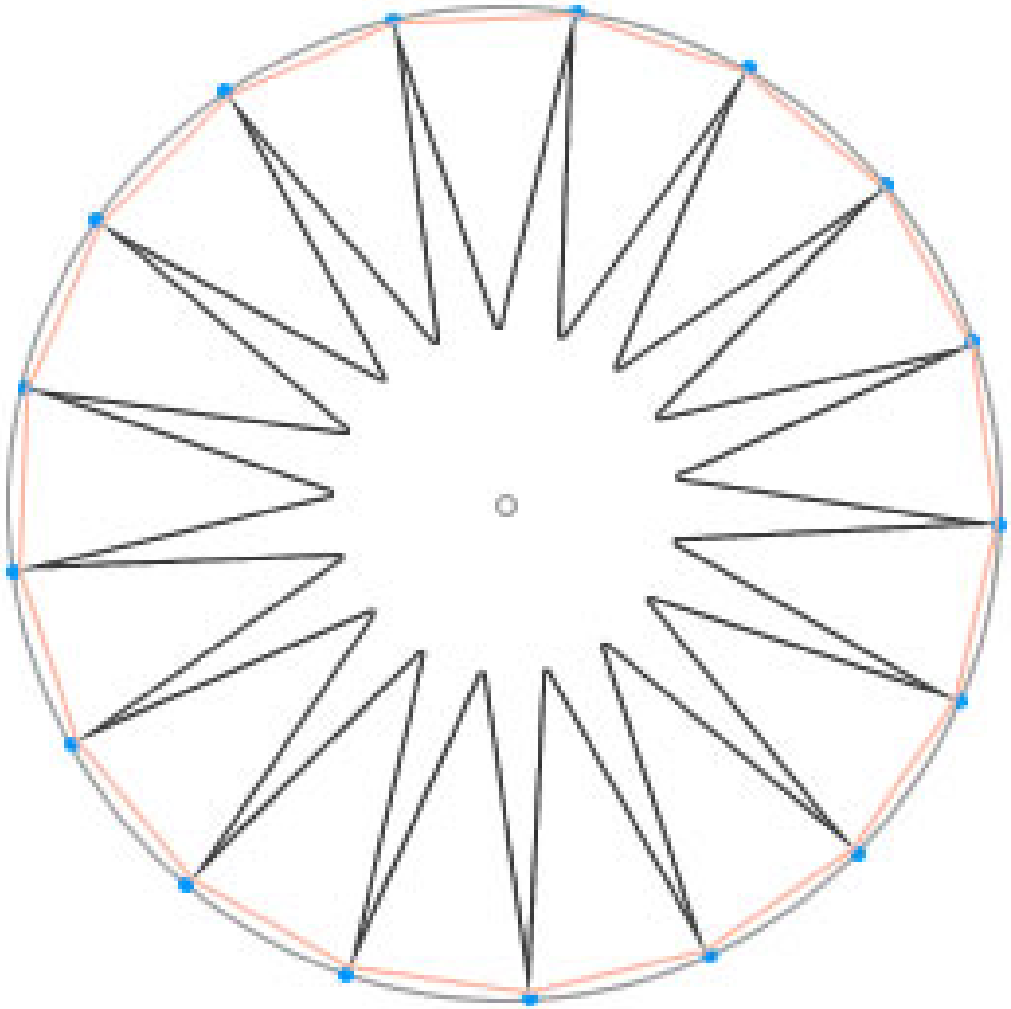
Spór Legendre'a z Gaussem o odkrycie MNK.
AMK był niewątpliwie pierwszy, który tę metode wykorzystał w opublikowanej pracy (1805), Gauss miał zwyczaj piszania do szuflady ...
Podobnie jak I. Newton, który toczył analogiczny spór z G. Leibnizem o odkrycie rachunku rózniczkowego i całkowego.
Pierwszy z publikacją był Leibniz *Nova methodus pro maximis et minimis*.
GL w sporze odwołał się do Royal Society (której prezesem w tym czasie był Newton ..).

Carl Friedrich Gauss



Carl Friedrich Gauss

Konstrukcja geometryczna 17-kąta foremnego:



Gauss' Regular "17-gon"

Carl Friedrich Gauss

W 1809 roku Gauss publikuje 'Teorię ruchu ciał niebieskich wokół słońca'. Pojawia się w niej MNK oparta na modelu probabilistycznym. Co zrobił Gauss ?

$$\begin{aligned} V &= ap + bq + cr + ds + etc. \\ V' &= a'p + b'q + c'r + d's + etc. \\ V &= a''p + b''q + c''r + d''s + etc. \end{aligned}$$

V, V', V'' - wartości oczekiwane odpowiedzi, a, b, c - zmienne niezależne, $p, q, r..$ -współczynniki.
Równania strukturalne dopasowywane do danych. M, M', M'' - obserwowane wartości zmiennej zależnej ($EM = V$ etc).

$$\Delta = V - M,\ \Delta' = V' - M',\ \Delta'' = V'' - M''\ etc.$$

błędy pomiaru

Prawdopodobieństwa błędów pomiaru

$$\phi(\Delta),\ \phi(\Delta'),\ \phi(\Delta'')\ etc.$$

Wartości błedów znajdziemy maksymalizując

$$\Omega = \phi(\Delta) \times \phi(\Delta') \times \phi(\Delta'')\ etc.$$

O ϕ przyjął postulaty:

- $\mathrm{argmax}\phi(s) = 0$;
- $\phi(s) = \phi(-s)$;
- Gdyby $V = V' = V'' = .. = p$ to wartościa maksymalizującą Ω byłoby $\hat{p} = k^{-1}(M + M' + M'' + etc)$, gdzie k jest liczbą obserwacji.

Kołowość 3-ciego postulatu: zakładamy coś, co mamy pokazać !

Przy tych założeniach Gauss pokazał, że

$$\phi(\Delta) = \frac{h}{\sqrt{\pi}} e^{-h^2\Delta^2}$$

i że przyjęcie takiej krzywej błedów prowadzi do estymacji metodą najmniejszych kwadratów.
Estymator NW dla modelu regresji liniowej z błędami normalnymi jest estymatorem MNK.

$$\mathcal{L} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(Y_i - x_i'\beta)^2}{2\sigma^2}) = \frac{1}{\sqrt{2\pi\sigma^2}^n} \exp(-\frac{\sum_{i=1}^{n}(Y_i - x_i'\beta)^2}{2\sigma^2})$$

Postać estymatora MNK metodą Gaussa

FG: inne rozwiązanie nie wykorzystujące równań normalnych. Napisał model liniowy

$$y = X\beta + \varepsilon$$

w postaci

$$X'X\beta = X'y - X'\varepsilon = X'y + z,$$

gdzie $z = -X'\varepsilon$.
FG szukał estymatora b parametru β w postaci

$$\beta = b + Qz.$$

Zatem

$$QX'X\beta = QX'y + Qz = QX'y + \beta - b$$

ma zachodzić dla dowolnego β zatem

$$QX'X = I_p \qquad Q = (X'X)^{-1} \qquad b = QX'y$$

Analiza rezyduów

FG udowodnił i stwierdził, że jest to ciekawe, iż

$$\mathrm{Var}(\varepsilon'\varepsilon/n) = 2\sigma^4/n$$

$$\mathrm{Var}(e'e/n) = 2\sigma^4/(n-p),$$

gdzie p jest liczbą predyktorów. Wniosek: $e'e$ może być traktowana jaka suma kwadratów $n-p$ niezależnych zmiennych losowych.
Komentarz: niestety nie zauważył, że

$$\varepsilon'\varepsilon = e'e + (b-\beta)'(X'X)(b-\beta)$$

i że oba składniki są niezależne, bo to by dało mu postać testu F odkrytego przez Fishera 100 lat później ($H_0 : \beta = 0$. Przy H_0

$$\frac{p^{-1}b'(X'X)b}{(n-p)^{-1}e'e} \sim F_{p,n-p}$$

CTG Laplace'a

Równolegle do prac nad podejściem bayesowskim (inverse probability) Laplace zajmował się CTG.
1776: CTG dla zmiennych losowych o rozkładzie prostokątnym
1781: dla zmiennych losowych o gęstości kawałkami ciągłej.
1810: przypadek ogólny. .....

Laplace po zobaczeniu wyniku Gaussa zrozumiał, że w oparciu o CTG ma znacznie lepsze uzasdnienie krzywej błędów, czego przedtem nie widział. Dodał to jako uzupełnienie do CTG w swojej pracy z 1810.
Spór o pierwszeństwo miedzy Legendre'm i Gaussem.
Legendre mowił o MNK jako o naszej (tj Gaussa i jego) metodzie, Gauss twierdzil, że używal tej metody od 1795 roku i mówił swoim współpracownikom o niej przed 1805 rokiem.

Twierdzenie Gaussa-Markowa : Laplace

L. rozpatrywał sytuację

$$\varepsilon^{(i)} = p^{(i)}z - \alpha^{(i)}, \quad i = 1, \ldots, n$$

$\varepsilon^{(i)}$ błędy aproksymacji $y_i = \alpha^{(i)}$ funkcją liniową $p^{(i)}z$ ($p^{(i)} = x_i$).
Rozumowanie było oparte na prostej obserwacji: gdyby dla pewnych wag $m^{(i)}$ mielibyśmy $\sum m^{(i)}\varepsilon^{(i)} = 0$ wtedy

$$\sum m^{(i)}\varepsilon^{(i)} = \sum m^{(i)}p^{(i)}z - \sum m^{(i)}\alpha^{(i)}$$

i

$$z = \frac{\sum m^{(i)}\alpha^{(i)}}{\sum m^{(i)}p^{(i)}}$$

Oczywiście z reguły $\sum m^{(i)}\varepsilon^{(i)} \neq 0$. Wtedy modyfikacja rozumowania daje

$$z = \frac{\sum m^{(i)}\alpha^{(i)}}{\sum m^{(i)}p^{(i)}} + u,$$

gdzie

$$u = \frac{\sum m^{(i)}\varepsilon^{(i)}}{\sum m^{(i)}p^{(i)}}.$$

Uogólniając swoje CTG pokazał w 1811 roku, że u ma rozkład asymptotycznie normalny oraz estymator minimalizujący oczekiwany błąd to

$$z = \frac{\sum p^{(i)}\alpha^{(i)}}{\sum p^{(i)2}},$$

czyli estymator najmniejszych kwadratów.
W 12 lat później Gauss udowodnił to samo dla błędu średniokwadratowego i zauważyl, że wystarczy tylko skończony drugi moment błedów. Twierdzenie Gaussa-Markowa.