Historia Rachunku Prawdopodobieństwa i Statystyki
WYKŁAD VI: Galton, Edgeworth i Karl Pearson.

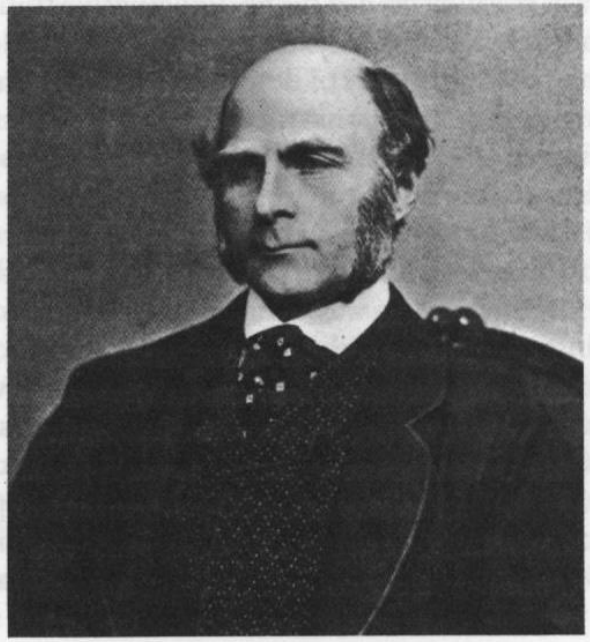
MiNI PW

Galton, Edgeworth i Pearson

Trzej badacze, którzy razem, korzystając wzajemnie ze swoich wyników osiągnęli więcej, niż każdy z nich osiągnąłby osobno (efekt synergii).

- Francis Galton: miał bardzo dobre pomysły, ale nie miał dobrych podstaw matematycznych i nie umiał uogólnić swych częsciowych wyników;
- Francis Edgeworth: dobry matematyk, umiał wyciągnąc wnioski ogólne z wyników częściowych, mial problemy z przekazaniem ich szerzej i nie umiał zebrać danych, aby swe ogólne wnioski potwierdzić;
- Karl Pearson: nie miał zdolności Galtona i głebokości Edgewortha, ale cechowały go ambicja i upór, które pozwoliły mu, razem z Yule'm na stworzenie związanej z odkryciami poprzedników spójnej metodologii i 'sprzedanie' jej światu.

Francis Galton (1822-1911)



Francis Galton (1822–1911)

Francis Galton (1822-1911)

FG był lekarzem w Cambridge, który po odziedziczeniu spadku po ojcu w wieku 22 lat przerwal pracę zarobkową i zajął się działalnoscią naukową.
Główne zainteresowania:

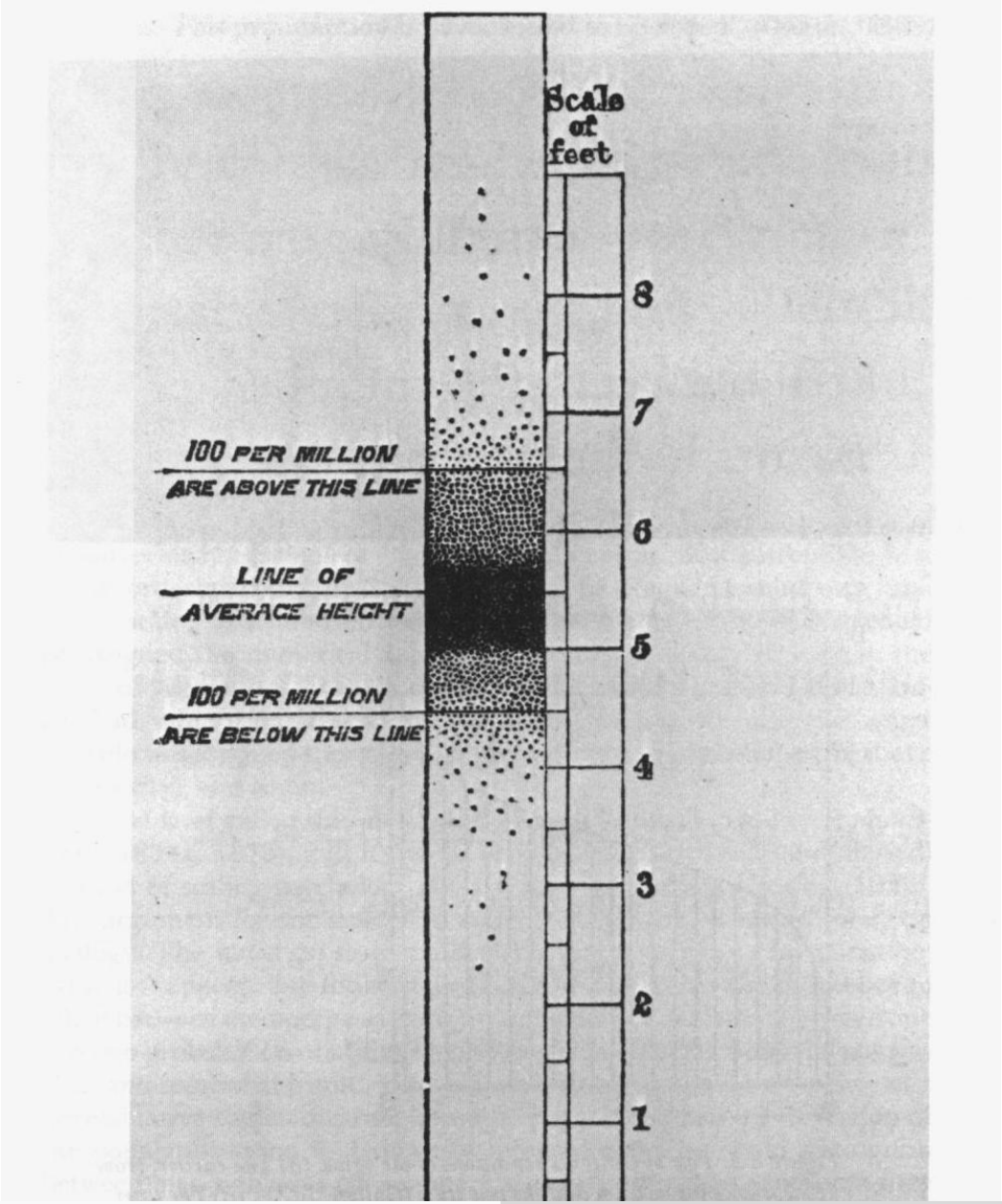
- eksploracja Afryki 1850–1852 (złoty medal Królewskiego Towarzystwa Geograficznego w 1853);
- meteorologia (sporządzenie map meteorologicznych);
- psychologia, antropologia, socjologia, edukacja...
- głowne zainteresowania po 1865 dotyczyły dziedziczności.

Być może jeden z ostatnich 'naukowców gentelmenów' *gentlemen scientists* (Stigler), niezależnych finansowo naukowców nie związanych z żadną instytucją naukową.
Stephen Wolfram ?
Merytorycznie był potomkiem Queteleta, podobnie jak on był zafascynowany zbieraniem i interpretacja danych.

Główne osiągnięcia:

- Narzędzia eksploracyjnej analizy danych (wykres pudełkowy, wykres kwantylowy (*ogive*)
- korelacja;
- regresja
- prawa dziedziczności, problemy dziedziczenia zdolności;
- quincunx (deska Galtona);
- biometria: linie papilarne są ważnym narzędziem identyfikacji (twórca Antropometric Laboratory i autor książki na ten temat);
- .....

Wykres pudełkowy



Ogiwa: protoptyp funkcji kwantylowej

FG przedstawiał rozkład w próbie na wykresie zależności wartości cechy od jej rangi $/n$ ($X_{i:n}$ vs i).
Zbadał jej kształt dla populacji normalnej i nazwał tę krzywą ogiwą

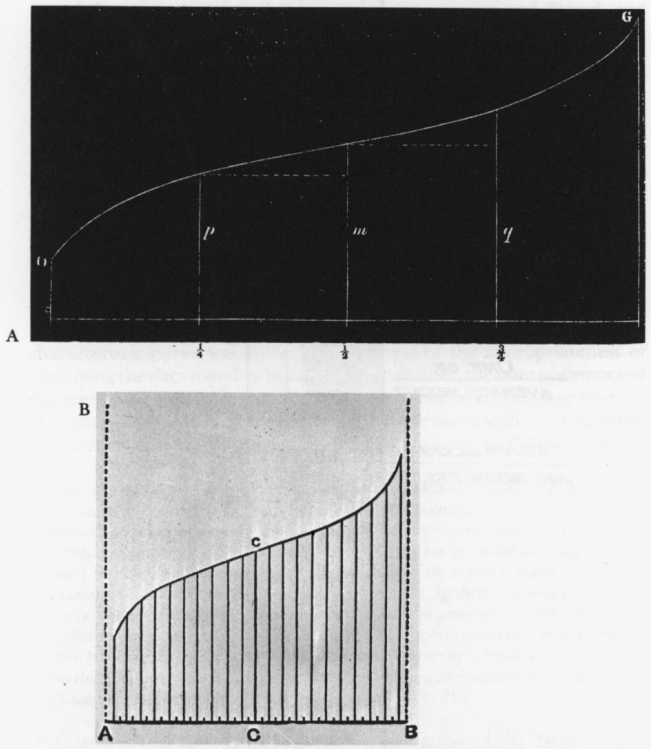


Figure 8.2. Two of Galton's renditions of the ogive. (A) The earlier, from 1875, shows the median and the quartiles (Galton, 1875). (B) The later drawing shows spikes with heights representing 21 equally spaced ideal data values (Galton, 1883, p. 51).

Ogiwa

Quincunx - deska Galtona

https://www.youtube.com/watch?v=6YDHBFVIvIs

Problem dziedziczenia i oceny zdolności

Miara ilościowa stopnia zdolności:
Porządkował ludzi w grupie względem zdolności i i-tej randze przyporządkowywał $\Phi^{-1}((n+i-1/2)/2n)$.
Metody analizy wzrostu i wagi przenosił na wszystkie inne cechy ...
Opublikował książke *Hereditary Genius*, w której starał sie udowodnić, na podstawie zebranych danych o rodzinach polityków, naukowców, poetów itp., że talent jest dziedziczny.

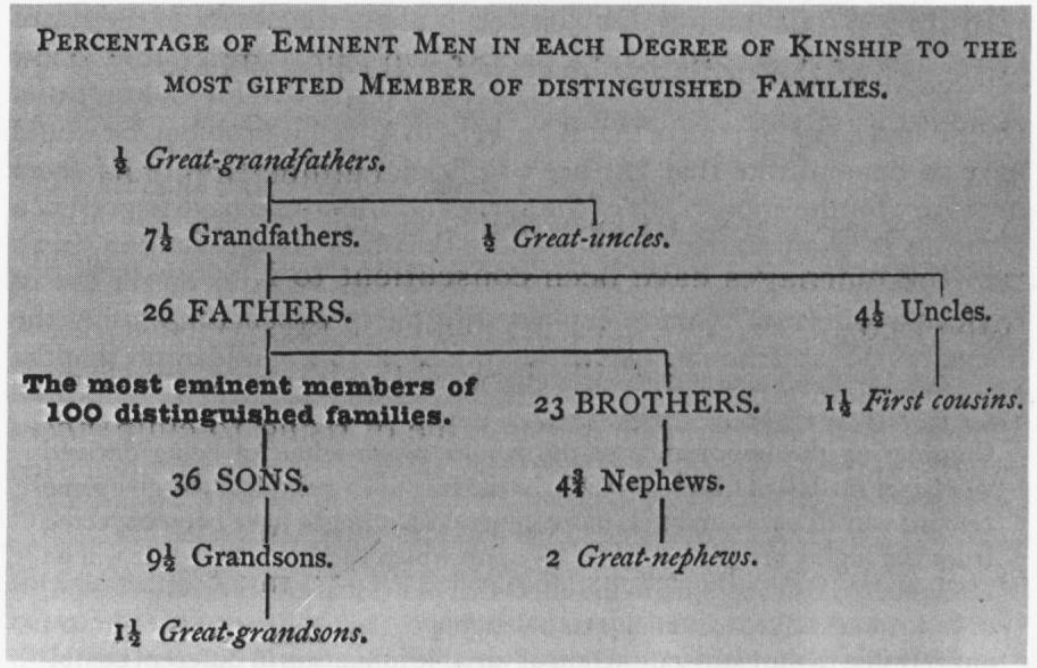


Figure 8.3. Galton's illustration of how the percentage of eminent men declines as the distance from the most eminent man in the family increases. Thus in 36 percent of the cases the son of the most eminent man was also eminent, whereas only 9.5 percent of the grandsons achieved eminence. (From Galton, 1869, p. 83.)

Regresja w kierunku średniej

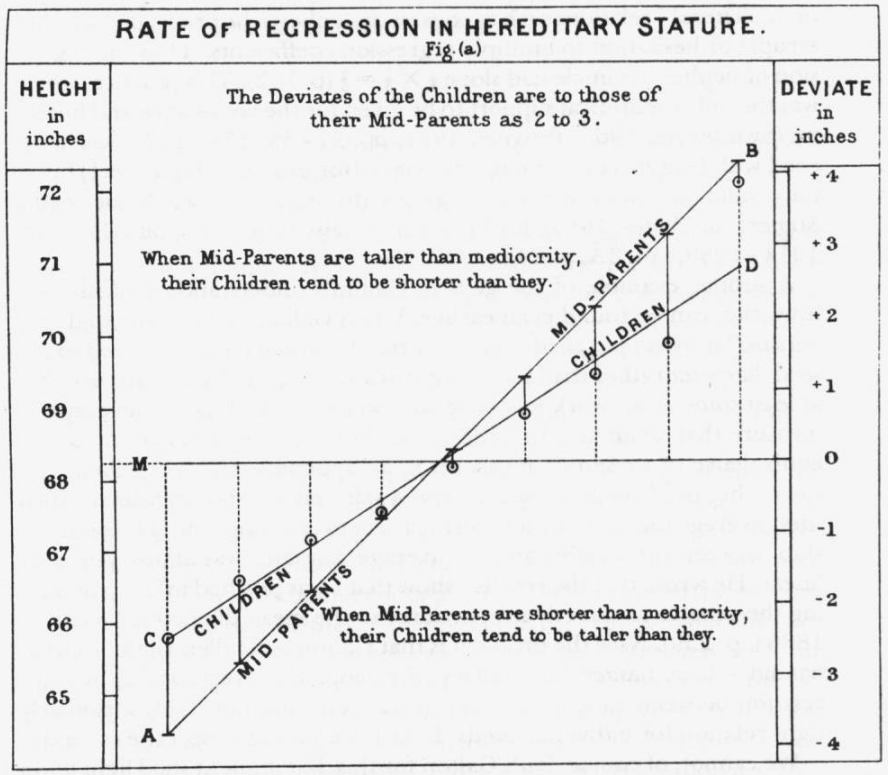


Figure 8.8. Galton's graphical illustration of regression; the circles give the average heights for groups of children whose midparental heights can be read from the line AB. The difference between the line CD (drawn by eye to approximate the circles) and AB represents regression toward mediocrity. (From Galton, 1886a.)

Kółka odpowiadają średniej wysokości dzieci w grupach dla ustalonej wysokości rodziców (linia AB). Linia CD nie jest prostą MNK i nie uwzględnia ostatniej obserwacji odstającej

Regresja w kierunku średniej - trzy wyjaśnienia

- wyjaśnienie intuicyjne;
- własność rozkładu normalnego;
- wyjaśnienie geometryczne.

Regresja w kierunku średniej - trzy wyjaśnienia

Intuicja

X- wynik na pierwszym teście, Y: wynik na następnym teście.
Bardzo dobry wynik X jest rezultatem dobrego przygotowania oraz (losowych) okoliczności, które mogą nie powtórzyć się następnym razem. W przypadku mniej sprzyjających okoliczności wynik Y nie bedzie już tak dobry jak wynik X relatywnie średniego wyniku.

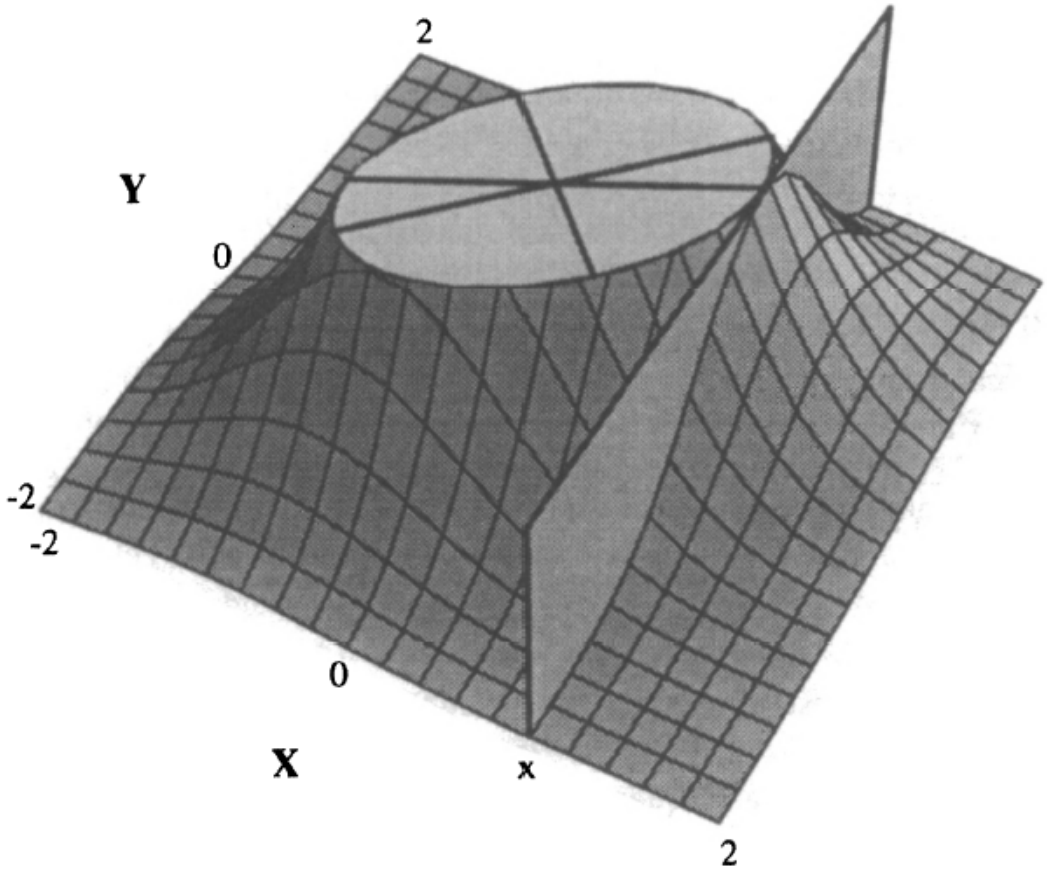
Własność rozkładu normalnego

Dla dwuwymiarowego rozkładu normalnego
$N_2(EX, EY, \sigma_X, \sigma_Y, \rho) \sim (X, Y)$

$$\frac{E(Y|X) - EY}{\sigma_Y} = \rho\Big(\frac{X - EX}{\sigma_X}\Big)$$

Galton otrzymał, że nachylenie prostej dla potomków = (2/3)x nachylenie prostej dla rodzicow i stosował stałą 2/3 w analizie innych charakterystyk, np zdolności artystycznych ...

Regresja w kierunku średniej - wyjaśnienie geometryczne



Gęstość normalna $N(0, 0, , 1, 1, 0.5)$. Na rysunku poziomica gęstości (elipsa wraz z osiami $y = x$ i $y = -x$) oraz linia warunkowej wartości oczekiwanej $E(Y|X = x)$. Ta ostatnia ma mniejsze nachylenie niż poziomica $y = x$.

Regression fallacy

Dla pewnego procesu w pewnym momencie obserwujemy dużą wartość odstającą, stosujemy interwencję i nastepna wartość procesu jest już mniejsza.
Regression fallacy: przekonanie, że jedynym wytłumaczeniem zmniejszenia się wartości procesu jest nasza interwencja ...
X_1, X_2 niezależne wartości pochodzące z rozkładu jednomodalnego.
Jeśli $X_1 = x$ duże to $P(X_2 > X_1 | X_1 = x) = P(X_2 > x_1)$ - małe.
Regresja w kierunku średniej.
Secrist *The Triumph of Mediocrity in Business* (1933) Wyniki bardzo dobrych firm w warunkach gospodarki rynkowej pogarszają się w czasie, bo następuje regresja w kierunku średniej..
Krytyka Hotellinga

Problem Galtona

Nie umiał przez długi czas poradzić sobie z problemem:
Jak to możliwe, że dla cechy, której rozklad w populacji jest normalny (np. dla wzrostu), a więc powinna być, na podstawie warunków zapewniających CTG (warunki Laplace'a) wynikiem działania wielu cech o mniej wiecej takiej samej sile, a jednocześnie cecha rodzicielska się dziedziczy , czyli jest dominująca ..
Galton doszedł do wniosku (1875), ze CTG musi zachodzić również przy innych załozeniach niż warunki Laplace'a.
Przedstawił koncepcję modyfikacji Quincunx'a, która pokazywała to zjawisko

CTG przy innych warunkach niż dla ciągu iid

Galton: na wielkość owocu ma wpływ nasłonecznienie (małe, średnie i duże). Istotne jest to, ze dwie sytuacje skrajne zdarzają sie znacznie rzadziej niż sytuacja typowa (nasłonecznienie średnie)

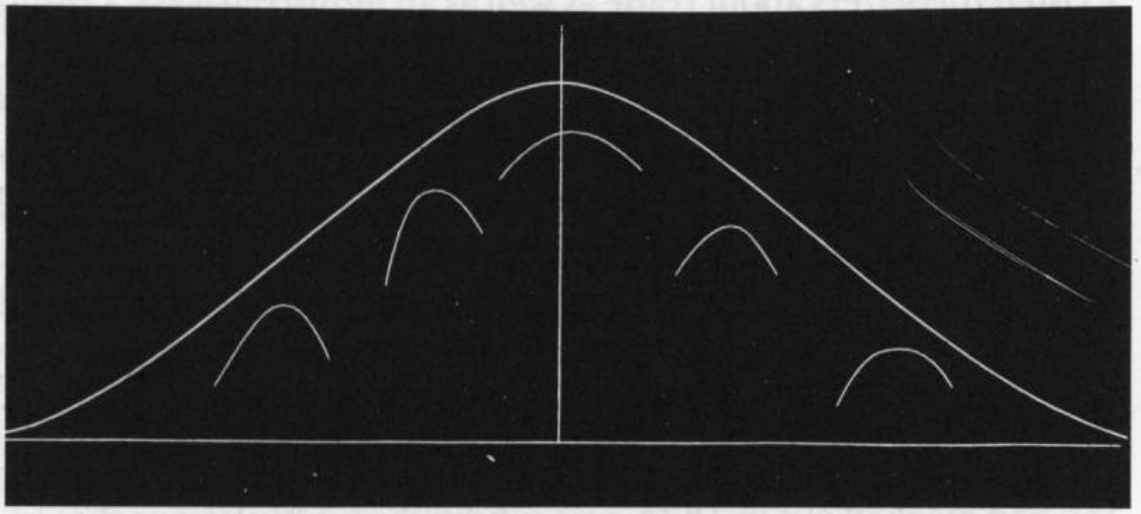


Figure 9.1. Edgeworth's 1885 illustration of Galton's 1875 insight. The smaller curves represent the distributions of the sizes of pears from specific parts of a garden (e.g., near a particular wall), and the large curve gives the distribution for the whole garden, thus showing the normal curve as a mixture of normal components. (From Edgeworth, 1885b.)

..

Francis Ysidro Edgworth (1845-1926)

Potomek bogatej rodziny irlandzkiej, ekonomista, prawnik i matematyk.
Rozwinięcie Edgwortha:
X_i- iid o średniej μ i wariancji σ^2.

$$Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

$$F_n(x) = P(Y_n \leqslant x)$$

$$F_n(x) = \Phi(x) - \frac{1}{6n^{1/2}} \lambda_3 \Phi^{(3)}(x) + \frac{1}{n} \ldots$$

Oszacowanie Berry-Essena okazało się bardziej praktyczne ...

$$\sup_t |F_n(t) - \Phi(t)| \leqslant L \frac{E|X_1|^3}{\sigma^3 \sqrt{n}}$$

Karl Pearson (1857 -1936)

Z wykształcenia prawnik o szerokich zainteresowaniach, od 1892 roku, pod wpływem Galtona, zajał się statystyką. Zajmował się również germanistyką (!), socjologią i etnografią.
Pisał się Carl Pearson lub później, Karl Pearson.
Książki:

- *Regression, Heredity, and Panmixia*,
- *On the general theory of skew correlation and nonlinear regression*
- *On the theory of contingency*

Karl Pearson (1857 -1936)

Autor 650 prac naukowych ('nigdy nie odbieram telefonów i nigdy nie chodzę na zebrania ..').
Był profesorem w University College w Londynie.
Założyciel jednego z najważniejszych pism biometrycznych/statystycznych *Biometrika* (1902) (i *Annals of Eugenics*(1925)[1] i pierwszego zakładu statystyki w University College (Londyn) (Departament of Applied Statistics - 1911)
Uczeń i biograf Galtona.

[1] Obecnie *Journal of Human Genetics*. Motto Darwina""I have no Faith in anything short of actual measurement and the rule of three"

Karl Pearson (1857 -1936)

Główne osiągnięcia:

- analiza dwuwymiarowego rozkładu normalnego;
- analiza rozkładów skośnych i rodzina rozkładów Pearsona;
- test chi-kwadrat zgodności rozkładów;
- analiza i modelowanie tablic kontyngencji;
- metoda momentów.

Metoda momentów

$X \sim P_{\theta_0}$, gdzie $P_\theta \in \{P_\theta\}_{\theta \in \Theta}$

$$\hat{m}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

Szukamy $\hat{\theta}_M$ takiego, że

$$E_{\hat{\theta}_M}(X^k) = \hat{m}_k$$

dla $k = 1, \ldots, L$, gdzie L-wymiar Θ.
Wady:

- może nie istnieć (rozwiązanie równań nie spełnia warunków nałożonych na rodzinę rozkładów;
- ma często dużą wariancję.

Ale często jest to najprostszy sposób konstrukcji estymatora ..

Rodzina rozkładów Pearsona

Rodzina gęstości spełniających równanie

$$\frac{d \ln p(x)}{dx} = \frac{x + a}{b_1 + b_2 x + b_3 x^2}$$

b_1- skośność μ_3/σ^3,
b_2-kurtoza $\mu_4/\sigma^4 - 3$;
b_3 -$\mu_5\mu_3/\sigma^8$.
Dopasowywane metodą momentów.

Test zgodności chi-kwadrat

$(X_1, \ldots, X_k)$ obserwacja z rozkładu wielomianowego $\mathrm{Mult}(n, p_1, \ldots, p_k)$ i

$$e_i = X_i - np_i$$

róznice między wartościami oczekiwanymi liczby obserwacji w komórkach a X_i. Statystyka chi kwadrat

$$\chi^2 = \sum_{i=1}^{k} \frac{e_i^2}{np_i} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

i udowodnił, że ma asymptotycznie rozkład χ^2 z $k-1$ stopniami swobody.

Rozkład statystyki chi kwadrat dla $k = 2$

$$\chi^2 = \sum_{i=1}^{k} \frac{e_i^2}{np_i} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

dla $k = 2$ $X_2 = n - X_1$ i $np_2 = n(1 - p_1)$. Kładąc $p_1 = p$ mamy

$$\chi^2 = \frac{(X_1 - np)^2}{np} + \frac{(n - X_1 - (n - np))^2}{n(1 - p)}$$

$$\chi^2 = (X_1 - np)^2 \left(\frac{1}{np} + \frac{1}{n(1 - p)} \right) = \frac{(X_1 - np)^2}{np(1 - p)}$$

$$\chi^2 = \left(\frac{X_1 - np}{\sqrt{np(1 - p)}} \right)^2$$

Teza wynika z CTG.

Test zgodności chi-kwadrat z estymowanym parametrami

Co w sytuacji, gdy p_i nie są znane a jedynie znamy ich formę parametryczną $p_i = p_i(\theta)$?. Estymujemy wartość parametru i

$$\hat{p}_i = p_i(\hat{\theta})$$

KP założył błednie, ze zamiana p_i na $\hat{p}_i$ nie spowoduje zmiany rozkładu asymptotycznego.
Długa kontrowersja z R. Fisherem, który udowodnił, ze rozklad asymptotyczny to χ^2_{k-l}, gdzie l liczba restrykcji na parametry.

Korelacja a przyczynowość, korelacje pozorne

K. Pearson widział roznicę miedzy zwiazkiem przyczynowym a korelacją:

Up to 1889 men of science had thought only in terms of causation, in future they were to admit other working category, that of correlation, and thus open to quantitative analysis wide fields of medical, psychological and sociological research

.

Spurious correlations:

Such high corralations arise from common growth or decline with time, when interpreted as causal or semicausal relationships, are in our opinin perfectly **idle**, *indeed they are only too apt to be mischievous, and we shall reach nothing, or les then nothing- knighthoods-by the investigation of them*

.

Spór z G. Yule'm dotyczący analizy danych jakościowych

Karl Pearson (1857 – 1936), in 1890

George Udny Yule (1871 – 1951)

George Udny Yule (1871-1951)

Był fizykiem eksperymentalnym i pracował jako pomocnik (*demonstrator*) K. Pearsona w University College. Razem z nim zaczął zajmować się statystyką. Wkrótce zaczął prowadzić niezależną działalność naukową.
W 1912 roku zaczął pracować w Cambridge jako Lecturer in Statistics.
Autor popularnego podręcznika *Introduction to the Theory of Statistics* (1911)
Prekursor NLP: ksiązka *Statistical Study of Literary Vocabulary*.

Spór z G. Yule'm dotyczący analizy danych jakościowych

KP uważał, że należy tablice kontyngencji traktować jako wynik obserwacji wartości dwuwymiarowego rozkładu normalnego i szukał takiego rozkładu normalnego, dającego prawdopodobieństwa równe częstościom klatek przy równości rozkładów brzegowych.
współczynnik korelacji tetrachorycznej.

Spór z G. Yule'm dotyczący analizy danych jakościowych

G. Yule uważał, że pewne zmienne są z natury dyskretnie i to co obserwujemy w tablicy kontyngencji to wartości dwuwymiarowej zmiennej o wartościach dyskretnych.
Wprowdził iloraz szans (odds ratio) zdefiniowany dla rozkładu 2x2

$$\theta = \frac{P(X=1|Y=1)}{P(X=0|Y=1)} \Big/ \frac{P(X=1|Y=0)}{P(X=0|Y=0)} = \frac{\pi_{11}\pi_{00}}{\pi_{01}\pi_{10}}$$

$$\hat{\theta} = \frac{n_{00}n_{11}}{n_{01}n_{10}}.$$

Q Yule'a

$$Q = \frac{\hat{\theta}-1}{\hat{\theta}+1} = \frac{n_{00}n_{11} - n_{01}n_{10}}{n_{00}n_{11} + n_{01}n_{10}}.$$

Spór z G. Yule'm dotyczący analizy danych jakościowych

Pearson i Heron (1913) 150-stronicowy atak na Yule'a i jego miarę (w *Biometrika*):
If Mr Yule's views are acccepted, incomparable demage will be done to the growth of modern statistical theory ..
Zaatakowano *'half-baked ideas'* Yule'a i *'specious reasoning'* i pogląd, ze Yule powinien wycofać swoje koncepcję, jeśli chce zachować swoją reputację jako statystyk.
W zależności od danych jedno z konkurencyjnych podejść KP i BY ma rację bytu ...

Inne osiągnięcia Yule'a

Analiza regresji dla pauper data (dane Booth'a (1984)).

Table 10.1. Yule's "correlation table" showing the relationship between pauperism and prevalence of out-relief, using data from 1891.

	Percentage of population in receipt of relief								
Ratio	0–1	1–2	2–3	3–4	4–5	5–6	6–7	7–8	8–9
0–1	1	9	13	3	3	1	—	1	1
1–2	—	18	17	7	—	1	—	—	—
2–3	2	19.5	23.5	18	5	1	—	—	—
3–4	1	24	30	31	10	4	—	—	—
4–5	—	11	34	26	19	7	—	—	—
5–6	—	9	14	27	20	11	1	—	—
6–7	—	1	20	17	15	6	—	1	—
7–8	—	6	4	13	12	7	—	—	—
8–9	—	—	5	7	6	1	—	—	—
9–10	—	—	2	1	3	7	4	—	—
10–11	—	—	1	3	5	—	1	—	—
11–12	—	—	2	—	1	—	—	—	—
12–13	—	—	1	—	—	—	—	—	—
13–14	—	—	—	—	1	—	—	—	—
14–15	—	—	—	—	1	—	—	—	—
15–16	—	—	—	—	—	—	1	—	—
16–17	—	—	1	—	—	1	—	—	—

Source: Yule (1895, p. 609).

Note: "Ratio" is the number of out-paupers to one in-pauper; the data are for 580 poor law unions. Based on these data Yule computed the correlation coefficient as 0.388, with a probable error of 0.022.

out- door pauper - osoba dostająca zasiłek bez świadczenia pracy
in- door pauper - osoba dostająca zasiłek za pracę w przytułku.

'Ratio' oznacza frakcję liczby out-door to in-door (dostających zasiłek bez pracy do dostających zasiłek za pracę).
Pracę w in-door houses uważano za stygmatyzującą. Yule policzył korelację między średnimi w wierszach i Ratio: otrzymał 0.38 i uznał, wbrew wnioskom Booth'a, ze istnieje miedzy tymi dwiema wielkościami zależność dodatnia.

Analiza regresji: szacowanie $E(X|Y)$

$y = out/in$, x- średni procent ludzi na zasiłku

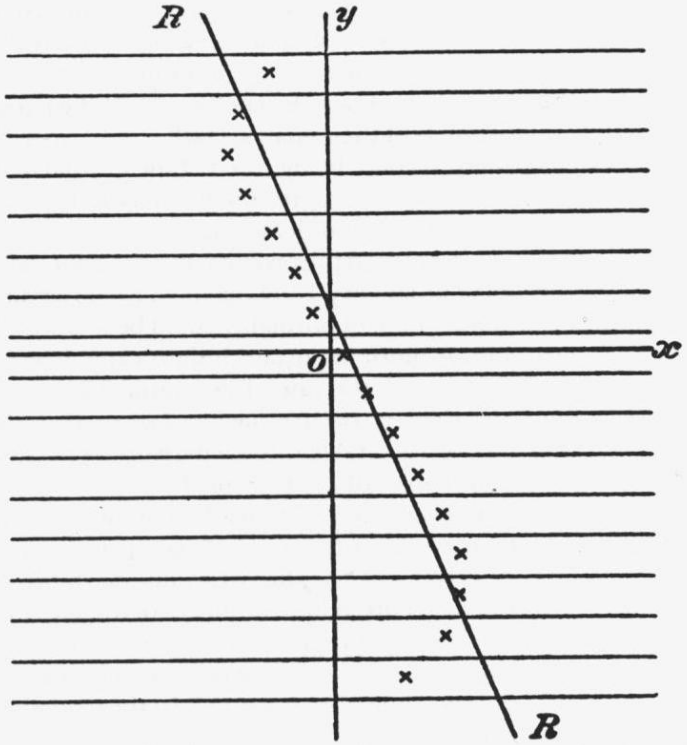


Figure 10.1. Yule's schematic drawing based on a correlation table. The x's represent means of the values within x-arrays (horizontal strips corresponding to rows of the table), and the line RR is the straight line closest (in a least squares sense) to the "curve of regression" given by the x's. (From Yule, 1897a,b.)

Yule stwierdził, że zwiekszanie pomocy bez obowiązku świadczenia pracy jest skorelowane z liczbą ludzi potrzebujących zasiłku..

Wyliczenia Yule'a

Yule dopasowywal prostą $a + by$ do średnich $\bar{x}_i = \sum_j x_{ij}/n_i$ i wyznaczył wspólczynniki minimalizując

$$\sum (\bar{x}_i - (a + by_i))^2 n_i$$

Stwierdził, ze jest to to samo, jakby a i b wyznaczył minimalizując

$$\sum_i \sum_j (x_{ij} - (a + by_i))^2,$$

czyli sprowadził problem do wyznaczenia prostej MNK dla danych (y_i, x_{ij}). Podstawowy fakt w analizie wariancji ANOVA.

Użycia i nadużycia testu chi kwadrat ..

Capricorn	63	(Lasker, Keres, Chiburdanidze)
Aquarius	79	(Spassky, Bronstein)
Pisces	101	(Fischer, Tarrasch, Geller, Larsen)
Aries	76	(Smyslov, Kasparov, Korchnoi, Portisch)
Taurus	77	(Miles, Nunn, Steinitz)
Gemini	67	(Petrosian, Karpov, Short, Euwe)
Cancer	54	(Morphy, Anderssen)
Leo	67	(Botvinnik)
Virgo	63	(Philidor)
Libra	69	(Rubinstein, Fine)
Scorpio	71	(Capablanca, Alekhine, Nimzowitsch, Tal)
Sagitarius	64	(Reshevsky, Pillsbury).

Przy H_0 $n\hat{p}_i = 851/12 = 70.917$. $.0025 < p = 0.0283 < 0.05$.
Przykład analizy post hoc.