STATISTICS FOR DATA ANALYSIS

EMPHASIS: METHODS THAT ARE INSENSITIVE TO HOW DATA OR ERRORS ARE DISTRIBUTED.

I. FITTING
- LINEAR LEAST SQUARES (ADVANTAGES & LIMITATIONS)
- MINIMUM ABSOLUTE DEVIATION; NONLINEAR FITTING
(LATER WE'LL TOUCH ON A NONPARAMETRIC ALTERNATIVE)

II. CONFIDENCE INTERVALS
- χ^2 (BRIEFLY)
- STATISTICAL BOOTSTRAP
→ INSTRUMENT FOCUS TEST EXAMPLE (FITTING W/UNKNOWN ERRORS)

Today

III. NONPARAMETRIC (DISTRIBUTION FREE) METHODS
→ EXAMPLE: MEDIAN VS. MEAN
- MEASURES OF CORRELATION
- NONPARAMETRIC FITTING
- KOLMOGOROV-SMIRNOV TEST
→ EXAMPLE: DISTRIBUTION OF NEARBY STARS

WE TAKE K-S TEST AS AN EXAMPLE OF FREQUENTIST HYPOTHESIS TESTING.

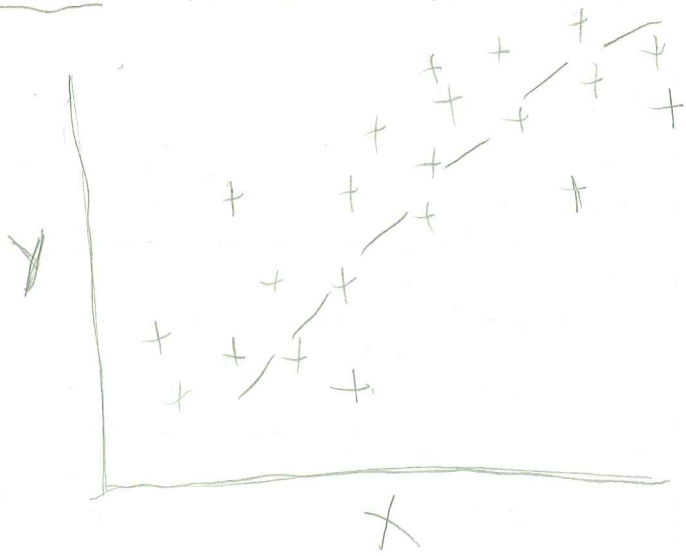
examples/Octave/statistical_analysis/descriptive_stats_examples/gauss_example.m & lorentzian_example.m

CORRELATION

A SUBSET OF STATISTICAL DEPENDENCE; APPLIES ONLY TO ORDINAL OR CONTINUOUS VARIABLES

GIVEN $\{X_1 \cdots X_N\}$ AND $\{Y_1 \cdots Y_N\}$

CORRELATION MEANS A MONOTONIC TREND IN Y VS. X



E.G.

$$Y = mX + b + \varepsilon$$

ε ← ADDITIVE NOISE

LINEAR CORRELATION COEFFICIENT:

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

NORMALIZATION STUFF →

r ASKS: WHEN $X_i > \bar{X}$, IS $Y_i > \bar{Y}$?

WHEN $X_i < \bar{X}$, IS $Y_i < \bar{Y}$?

IF $y = mx + b$ EXACTLY,

$$r = \begin{cases} 1 & m > 0 \\ -1 & m < 0 \end{cases}$$

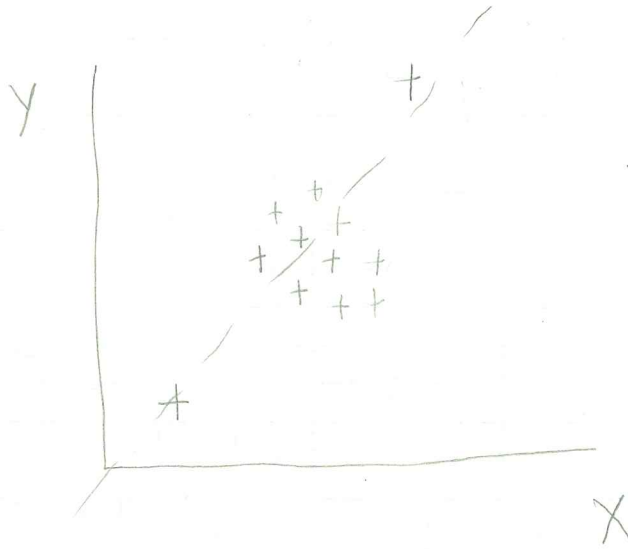
LINEAR CORRELATION COEFF. "r" IS DESIGNED FOR DATA (X,Y) WITH ADDITIVE GAUSSIAN NOISE.

PERFECT WORLD: $y' = mx' + b$

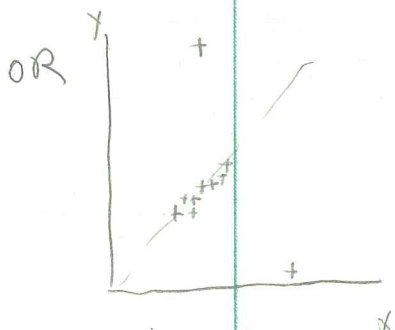
NOISED: $y = y' + \varepsilon_1$, $P(\varepsilon_1) \propto e^{-\varepsilon_1^2/2\sigma_1^2}$

$x = x' + \varepsilon_2$, ---- "

BUT "r" IS EASILY BROKEN BY A FEW "OUTLIERS:"



OR



THIS HAPPENS MOST OFTEN WITH NOISE THAT HAS FAT TAILS, E.G.

$$P(\varepsilon) \propto \frac{1}{1+x^2}$$

r SIMPLY "WEIGHTS" OUTLIERS TOO HEAVILY.

NONPARAMETRIC (RANK) CORRELATION STATISTICS (NR § 14.6)

GENERAL IDEA:

IF I SORT MY DATA (X,Y) BY ASCENDING X, HOW CLOSE IS THAT TO SORTING BY ASCENDING Y?

MOTIVATION:

MUCH LIKE THE MEDIAN, ALL THAT MATTERS IS RANK ORDER (NOT MAGNITUDE). ∴ INSENSITIVE TO OUTLIERS.

SPEARMAN RANK-ORDER CORRELATION COEFF.

① REPLACE (X_i, Y_i) WITH THEIR RANKS (R_i, S_i)

② "r_s" IS THE LINEAR CORRELATION COEFFICIENT USING (R, S) IN PLACE OF (X, Y):

$$r_s = \frac{\sum_{i=1}^{N}\left(R_i - \frac{N}{2}\right)\left(S_i - \frac{N}{2}\right)}{\sqrt{\sum_{i=1}^{N}\left(R_i - \frac{N}{2}\right)^2}\sqrt{\sum_{i=1}^{N}\left(S_i - \frac{N}{2}\right)^2}} \qquad -1 \le r_s \le 1.$$

(NOTE THAT $\frac{N}{2} = \bar{R} = \bar{S}$)

SPEARMAN IS FAST (LIMITED BY SORTING, $N \log N$ FOR HEAPSORT)

NR ROUTINE spear() GIVEN (X_i, Y_i) CALCULATES r_s AND SIGNIFICANCE.

KENDALL'S TAU (MORE NONPARAMETRIC / ROBUST / COMP. INTENSIVE THAN SPEARMAN)

- N DATA POINTS (X_i, Y_i)
- EXAMINE ALL PAIRS (X_i, Y_i) VS. (X_j, Y_j)
  (THERE ARE $\frac{1}{2}N[N-1]$ OF THEM!) $\Rightarrow$ $\mathcal{O}(N^2)$ COMPUTATIONS.
- CLASSIFY PAIRS AND COUNT HOW MANY ARE:

  N_c, CONCORDANT: $(X_j - X_i)$ SAME SIGN AS $(Y_j - Y_i)$

  N_D, DISCORDANT: $(X_j - X_i)$ OPPOSITE SIGN AS $(Y_j - Y_i)$

  $N_{=X}$, EXTRA-X: $X_i = X_j$ (BUT $Y_i \neq Y_j$)

  $N_{=Y}$, EXTRA-Y: $Y_i = Y_j$ (BUT $X_i \neq X_j$)

$$\tau \equiv \frac{N_c - N_D}{\sqrt{N_c + N_D + N_{=Y}} \cdot \sqrt{N_c + N_D + N_{=X}}}$$

$$-1 < \tau < 1$$

UNDER NULL HYPOTHESIS OF NO ASSOCIATION X VS. Y, τ IS APPROX. NORMALLY DISTRIBUTED, WITH $\bar{\tau} = 0$

AND $$\sigma_T = \sqrt{\frac{4N + 10}{9N(N-1)}} \rightarrow \sqrt{\frac{2}{3N}} \text{ FOR LARGE } N$$

NOTE: τ CAN BE SPED UP FOR BINNED DATA
(E.G. ORDINAL CONTINGENCY TABLE)

NR IMPLEMENTATIONS OF KENDALL'S τ:

kendl1() --- FOR 2D CONTINUOUS DATA
GIVEN (X_i, Y_i)
CALCULATES τ, SIGNIFICANCE

kendl2() --- FOR 2D ORDINAL (BINNED) DATA
GIVEN A TABLE T_{ij} OF COUNTS
CALCULATES τ, SIGNIFICANCE

BOTTOM LINE: (1) FOR UNBINNED DATA (X_i, Y_i), SPEARMAN IS FASTER. KENDALL OFFERS NO GREAT ADVANTAGE.
(2) FOR BINNED DATA, τ IS QUICKER AND MORE NATURAL. (IF # BINS IS MUCH SMALLER THAN #DATA POINTS).

A ROBUST ALTERNATIVE TO LINEAR LEAST SQUARES:

APPLICATION:

PORTER & KLIMCHUK 1995 ApJ 454, 499
OR
KLIMCHUK & PORTER 1995 NATURE 377, 131

FITTING $Y_i' = m X_i' + b$

WE MEASURE $Y_i = Y_i' + \varepsilon_i$, $X_i = X_i'$

GIVEN NO PRIOR KNOWLEDGE OF $P(\varepsilon)$ [BUT ASSUME INDEPENDENT OF X, Y] ⊛
→ EXCEPT WE ASSUME MEDIAN $(\varepsilon) = 0$.
(BUT YOU CAN BET $P(\varepsilon)$ IS NO WUSSY GASSIAN!)

APPROACH

OBVIOUSLY, WE EXPECT Y_i CORRELATED WITH X_i. CONSIDER:

① $Y_i^* \equiv Y_i - m^* X_i = b + \varepsilon_i + (m - m^*) X_i$

Y^* WILL BE UNCORRELATED (IN FACT INDEPENDENT) OF X IF $m^* = m$. THUS,

② MINIMIZE: $\tau(X, Y^*)$ W.R.T. m^*.

(THE RESULT IS $m^* \approx m$.)

③ LET $b^* = \text{MEDIAN}(Y^*)$ --- ASSUMES MEDIAN $(\varepsilon_i) = 0$.

ALTERNATIVELY, CONSIDER:

$$X_i = \frac{Y_i'}{m} - \frac{b}{m} = \frac{1}{m}(Y_i - \varepsilon_i - b)$$

LET $X_i^* \equiv X_i - \frac{1}{m^*} Y_i = \frac{-1}{m}(\varepsilon_i + b) + \left(\frac{1}{m} - \frac{1}{m^*}\right) Y_i$.

CAN WE MINIMIZE $\tau(X^*, Y)$ W.R.T. m^* AND GET THE SAME RESULT AS BEFORE?

NO: FOR $m = m^*$, $X_i^* = \frac{-1}{m}(\varepsilon_i + b)$. X^* IS STILL CORRELATED WITH Y BECAUSE ε IS CORRELATED WITH Y! (THOUGH ε IS UNCORRELATED WITH Y'!)

⊛ ASSUME $P(X', Y', \varepsilon) = P(X', Y') P(\varepsilon)$.

NONPARAMETRIC (RANK) CORRELATION IS A SUBSET OF "ROBUST ESTIMATION" METHODS (NR §15.7) *

VERY BROAD FIELD, EXPLODED c. 1960, STILL UNDER RESEARCH!

OBJECTIVE	A TRADITIONAL APPROACH	ROBUST METHOD(S)
CENTRAL TENDENCY OF $\{X_i\}$	MEAN	MEDIAN
COMPARE 2 DISTRIBUTIONS	MOMENTS	KOLMOGOROV-SMIRNOV STATISTIC
CORRELATION	LINEAR COEFF. "r"	NONPARAMETRIC CORRELATION E.G. KENDALL'S τ
FITTING TRENDS $y=mx+b$	LINEAR LEAST SQUARES	METHOD OF KLIMCHUK & PORTER; MIN ABS DEV
MULTI-PARAMETER FITTING $Y(x)$	NONLINEAR LEAST SQUARES	LOESS NONPARAMETRIC REGRESSION; MIN ABS DEV
ESTIMATING CONFIDENCE LEVELS	INTEGRATE PDF OF SOME STATISTIC, ASSUMING GAUSSIAN ADDITIVE NOISE IN THE DATA	MONTE CARLO (E.G. BOOTSTRAP) * NR §15.6

IN GENERAL --- ↑ FASTER - - - - ↑ MORE RELIABLE

GREATER STATISTICAL POWER (IF ALL ASSUMPTIONS MET)

(FEWER ASSUMPTIONS)

* SEE ALSO EFRON & TIBSHIRANI, SCIENCE 253, 390 (1991)

KOLMOGOROV-SMIRNOV TEST

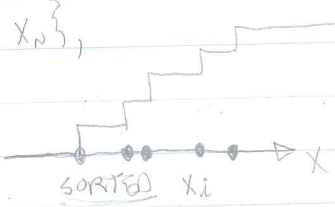
DEPENDS ON EVALUATING AND COMPARING CUMULATIVE DISTRIBUTION FUNCTIONS:

MY NOTATION FOR CUMULATIVE DIST.

$$C(x) = \int_{-\infty}^{x} P(x')\,dx'$$

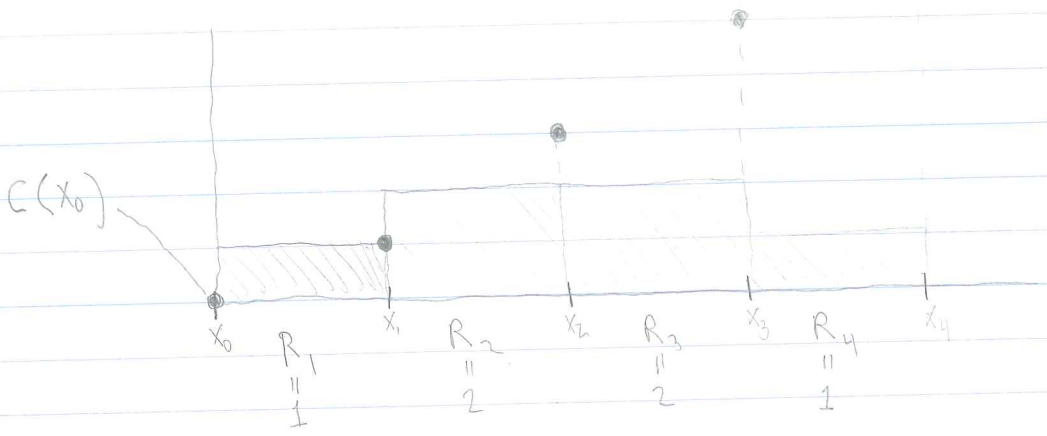
- WITH SAMPLED DATA $\{x_1 \cdots x_N\}$,

$$C(x) = \sum_{x_i \le x} \frac{1}{N}$$



WE HAVE C(x) AT 2N POINTS.

- WITH BINNED DATA, WE CAN ONLY EVALUATE C(x) ON THE BORDERS BETWEEN BINS:



THE K-S STATISTIC:

$$D = \text{MAX} \left| C(x) - C(y) \right|$$

- INVARIANT UNDER REPARAMETERIZATION OF x:

$$x \to \log x, \quad y \to \log y$$

NOTES ON K-S STATISTIC D:

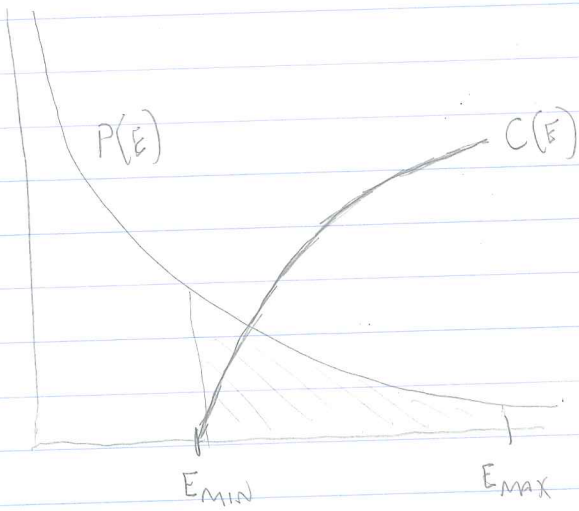
- VERY SENSITIVE TO MEDIAN $\bar{X}_m$.
- CONTAINS AN INTRINSIC NOTION OF FUNCTION WIDTH AND SHAPE THAT DOES NOT DEPEND ON THE EXISTENCE OF MOMENTS OF $P(x)$.
- WORKS FOR ANY $P(x)$ THAT CAN BE NORMALIZED

  I.E. $\int_{-\infty}^{\infty} P(x)\,dx$ CONVERGES.
- CAN BE EXTENDED TO <u>ANY</u> $P(x) \geq 0$ ON A RESTRICTED INTERVAL $[a,b]$ AS LONG AS

$$\int_a^b P(x)\,dx \quad \text{CONVERGES.}$$

E. G. POWER LAW DISTRIBUTION OF FLARE ENERGIES:

$$P(E) = E^{-\alpha}$$



$$C(E) = \int_{E_{MIN}}^{E} P(E')\,dE'$$

NULL HYPOTHESIS: $P(x) = P(y)$

K-S TEST:

$$\text{PROBABILITY}\ (D > \text{OBSERVED}) = Q_{KS}\left[D\left(\sqrt{N_e} + 0.12 + \frac{0.11}{\sqrt{N_e}}\right)\right]$$

WHERE: $Q_{KS}(\lambda) \equiv 2\sum_{j=1}^{\infty}(-1)^{j-1} e^{-2j^2\lambda^2}$ FOR $N_e \gg 1$

$\left(Q_{KS}(0) = 1, \quad Q_{KS}(\infty) = 0\right)$

★ — PLOT THIS! —

AND:

$$N_e \equiv \begin{cases} N & \text{IF COMPARING } \{X_1 \cdots X_N\} \text{ TO } P(x) \\ \dfrac{N_1 N_2}{N_1 + N_2} & \text{IF COMPARING } \{X_1 \cdots X_N\} \text{ TO } \{Y_1 \cdots Y_N\} \end{cases}$$

"EFFECTIVE NUMBER OF DATA POINTS"

FOR SMALL N_e, OR BINNED DATA, Q_{KS} MAY DIFFER.

POSSIBLE APPROACHES:

- DERIVE MORE GENERAL Q_{KS} ANALYTICALLY
- RUN A MONTE-CARLO TEST TO ESTIMATE STATISTICAL SIGNIFICANCE OF D