ME170b Lecture 8 3/15/24

Experimental Techniques

Last time:

> covariance

Today:

> Ch.9
>
> > correlation and covariance
>
> Review

Main Results on Covariance

$$q_i = q(x_i, y_i)$$

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}).$$

covariance — non zeros means they are not independent!

$$\sigma_q^{\,2} = \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x^{\,2} + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y^{\,2} + 2\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\sigma_{xy}$$

std of q given non independent samples!

$$\sigma_q \leqslant \left|\frac{\partial q}{\partial x}\right|\sigma_x + \left|\frac{\partial q}{\partial y}\right|\sigma_y$$

naive estimate is still upper bound!

$\boxed{\text{Schwarz inequality}}$

never greater even when x_i, y_i are not independent.

Coefficient of Linear Correlation

Q: Given a set of measurements (x1,y1) …. (xN,yN), how well do they support the hypothesis that x and y are <u>linearly related</u>?

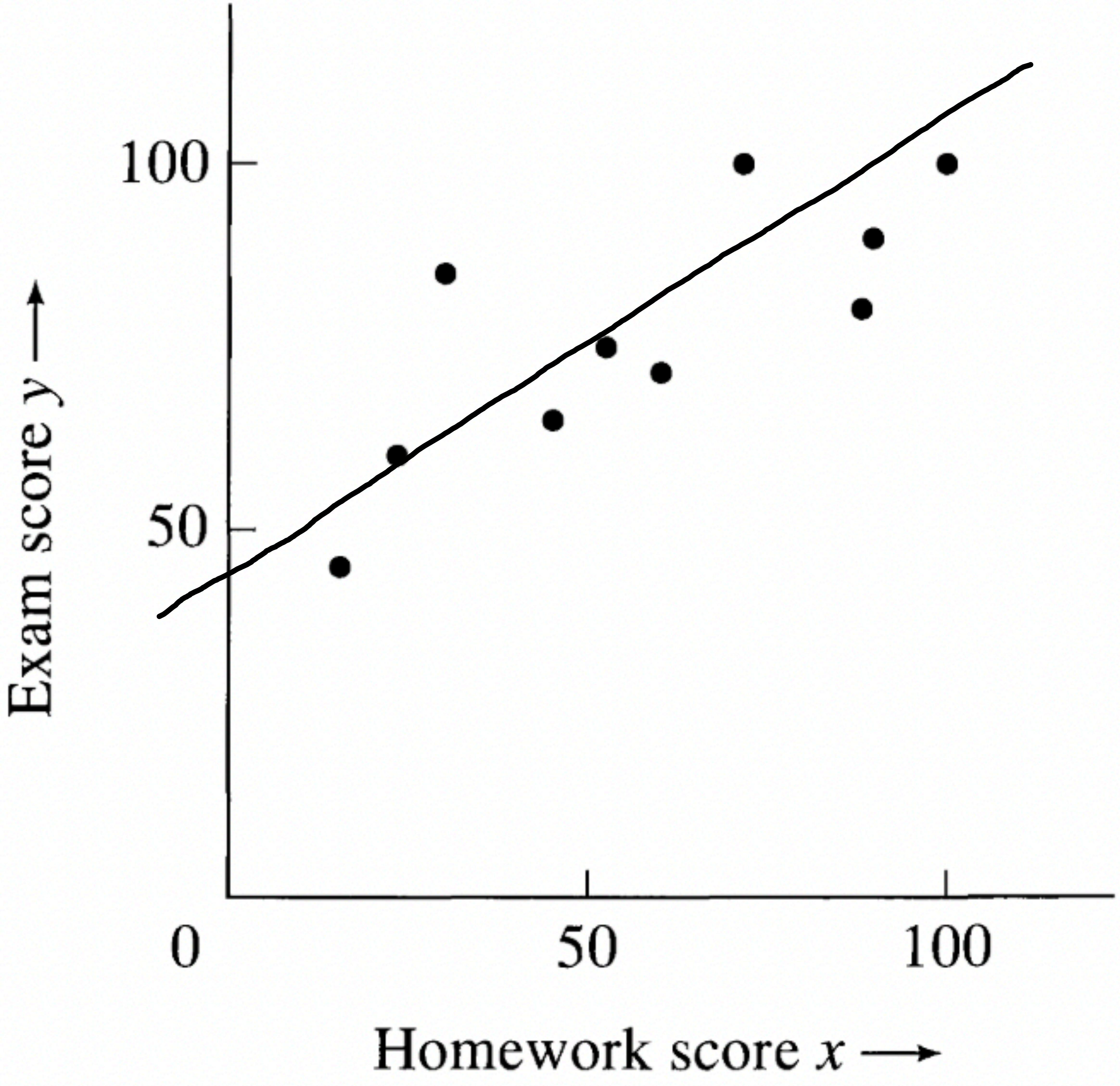
$$y = \underline{A} + \underline{B}x$$

Using the method of least squares, we can find the values of A and B for the line that best fits the points

If we already have a reliable estimate of the uncertainties in the measurements, we can see whether the measured points do lie reasonably close to the line

We don't have uncertainty measures (each measurement is different range), how can we determine how well our data fits

Example



Professor plots HW vs Exam score
Hypothesis: better HW -> better score

The professor hopes to show that high exam scores tend to be correlated with high homework scores, and vice versa

This kind of experiment has no uncertainties in the points; each student’s two scores are known exactly.

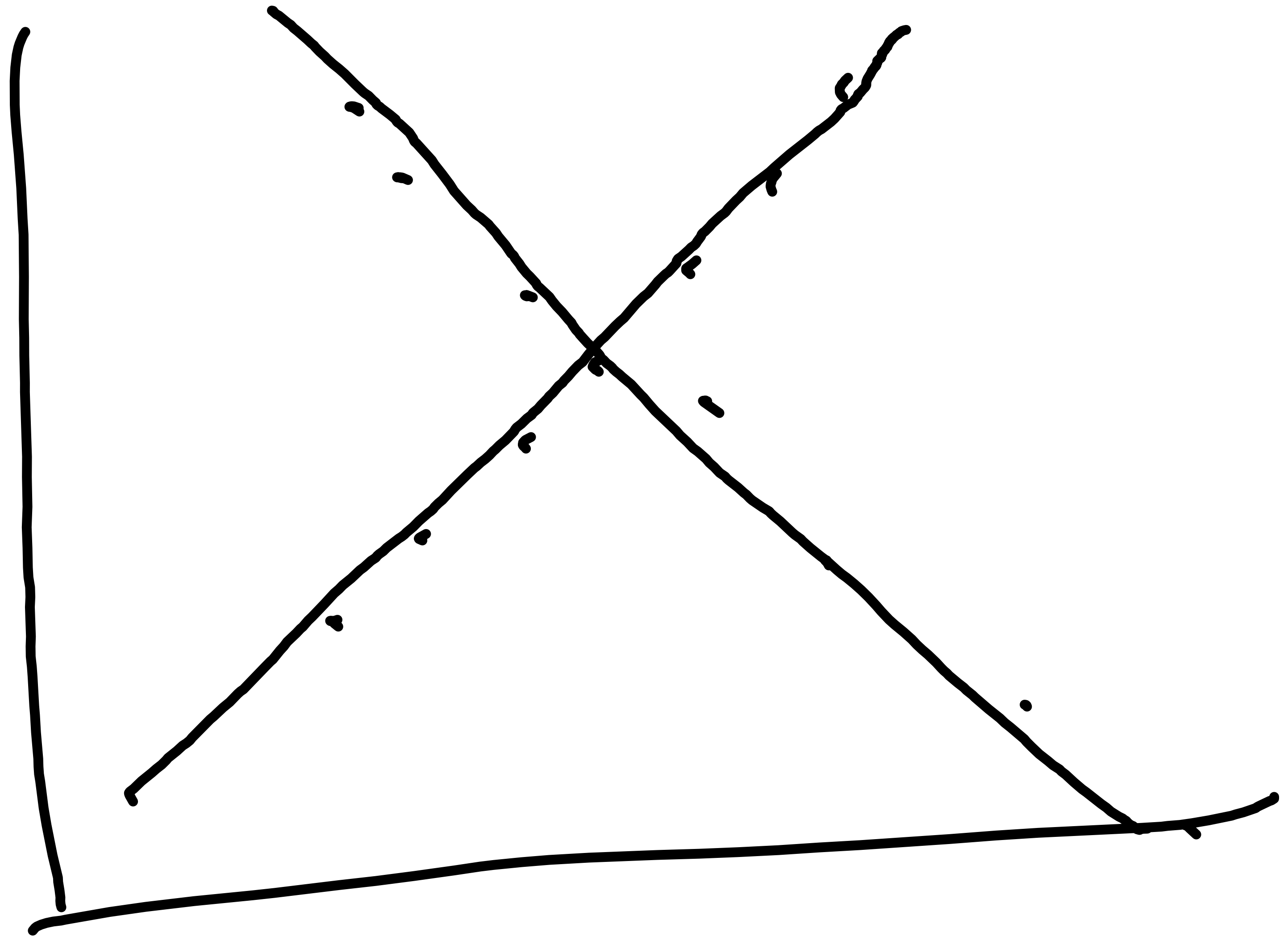
Really, what we want to understand the correlation

Correlation Coefficient

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

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

$$-1 \leqslant r \leqslant 1$$



Correlation Coefficient - perfect correlation

$$y_i = A + Bx_i$$

$$\bar{y} = A + B\bar{x}$$

$$y_i - \bar{y} = B(x_i - \bar{x})$$

$$r = \frac{B \sum (x_i - \bar{x})^2}{\sqrt{\sum (x_i - \bar{x})^2 B^2 \sum (x_i - \bar{x})^2}} = \frac{B}{|B|} = \pm 1$$

Correlation Coefficient revisit data

$$r = 0.8$$

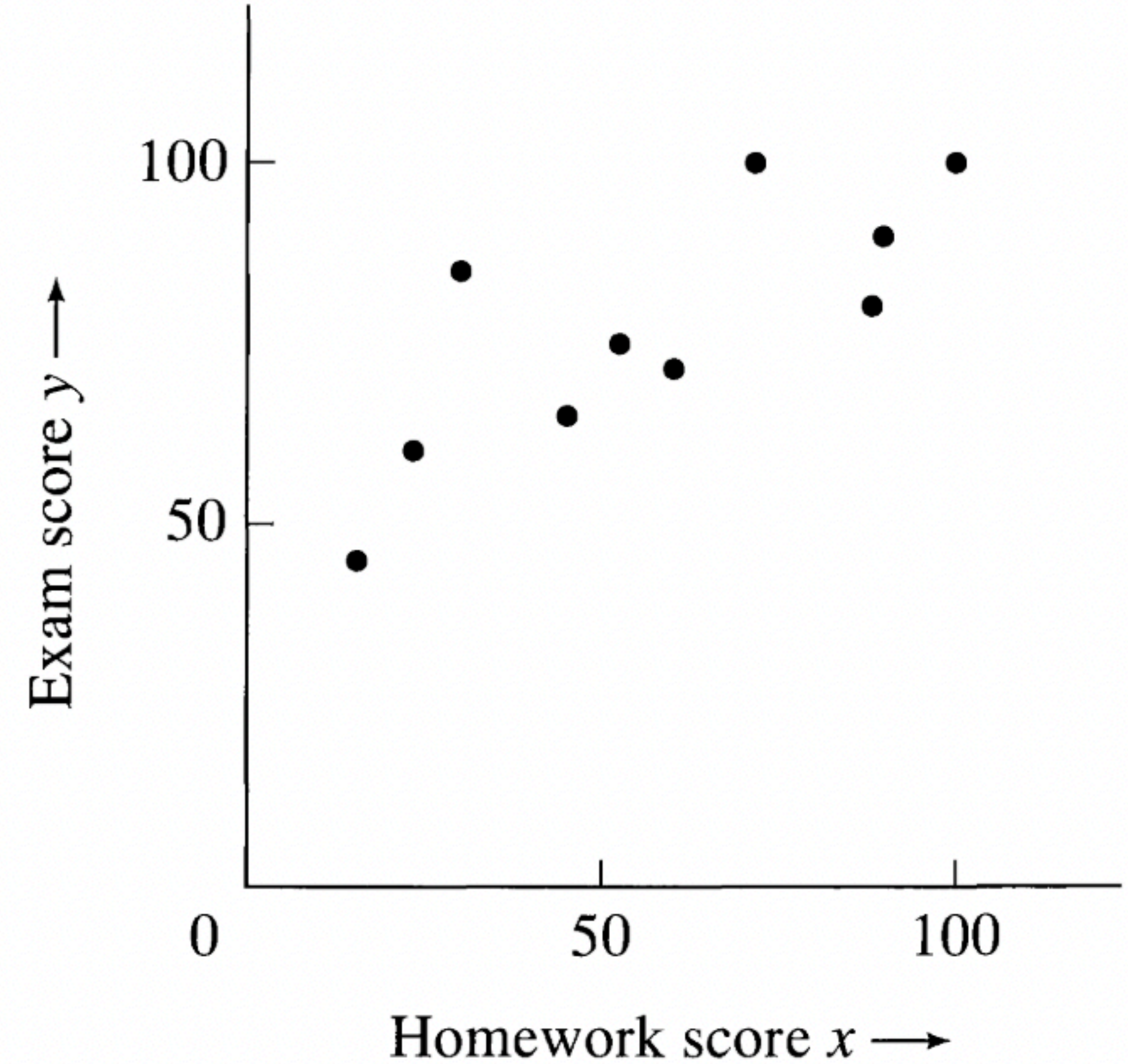


Table 9.3. Students' scores.

Student i	1	2	3	4	5	6	7	8	9	10
Homework x_i	90	60	45	100	15	23	52	30	71	88
Exam y_i	90	71	65	100	45	60	75	85	100	80

Find r → determine if significant? 95%.

Quantitative Significance of r

But how can we decide objectively what is good ‘r’?

Suppose the two variables x and y are in reality uncorrelated; that is, in the limit of infinitely many measurements, the correlation coefficient r would be zero.

We can calculate the probability that r will exceed any specific value:

$$Prob_N(|r| \geqslant r_o) \qquad Prob_N(|r| \geqslant 0.8)$$

Not straight forward calculation

Prob N measurements of two uncorrelated variables x and y would produce a correlation coefficient with

						r_o					
N	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
3	100	94	87	81	74	67	59	51	41	29	0
6	100	85	70	56	43	31	21	12	6	1	0
10	100	78	58	40	25	14	7	2	0.5		0
20	100	67	40	20	8	2	0.5	0.1			0
50	100	49	16	3	0.4						0

Given 10 random samples 0.5 %

↳ $r = 0.80$

***

Back to our original question

Q: Given a set of measurements (x1,y1) …. (xN,yN), how well do they support the hypothesis that x and y are linearly related?

1. we can calculate correlation coefficient
2. we can find the probability of observing r with uncorrelated
3. if the value is sufficiently small, we support our hypothesis!

“significant” if the probability of obtaining a coefficient r with |r| = |r| from uncorrelated variables is less than 5%. A correlation is sometimes called “highly significant” if the corresponding probability is less than 1%.

Main result: we have a quantitative measure of how improbable it is that they are uncorrelated.

Correlation Coefficient revisit data

$$r = 0.8$$

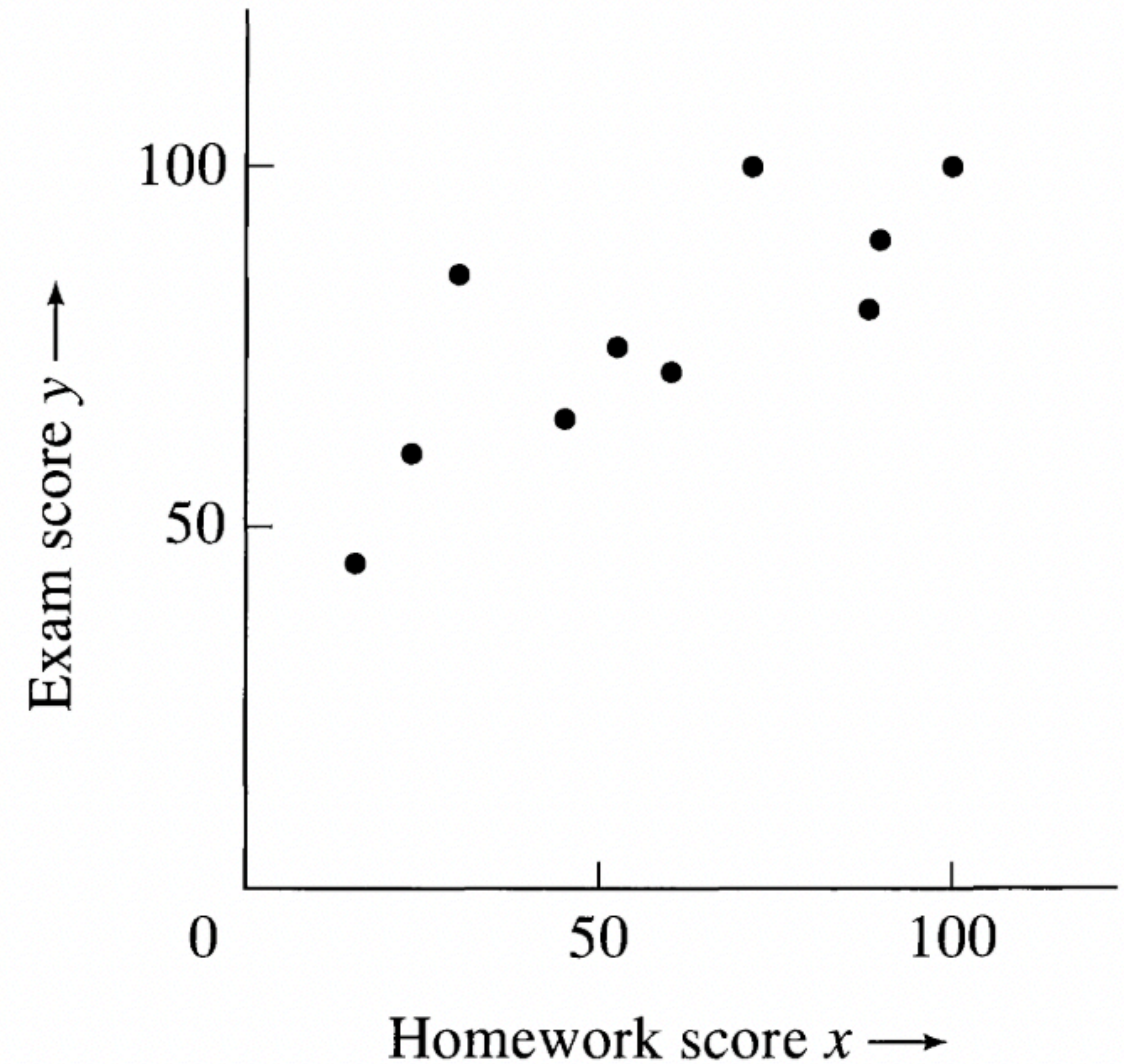


Table 9.3. Students' scores.

Student i	1	2	3	4	5	6	7	8	9	10
Homework x_i	90	60	45	100	15	23	52	30	71	88
Exam y_i	90	71	65	100	45	60	75	85	100	80

$\text{Prob}_{10}(|r| \geq 0.8)$ $N=10$

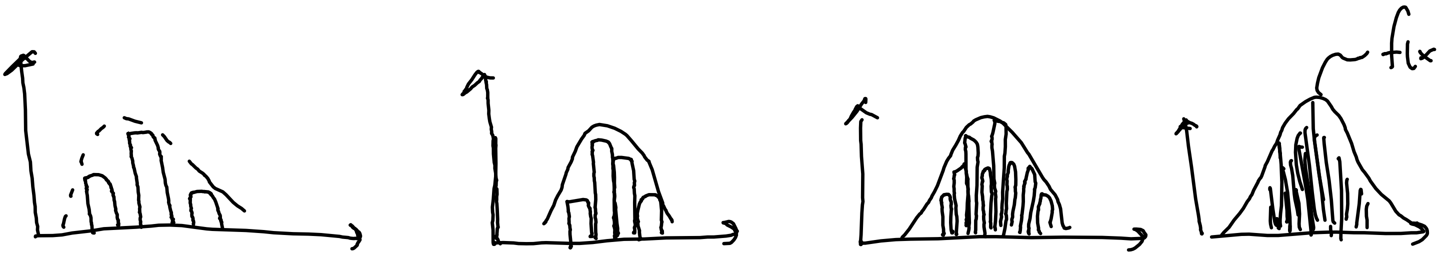
Exam - 2 (Final) 3/19 @ 3 - 6pm

Canvas will open @ 2:55 pm

close @ 6:05pm

- include any external code/excel files.
- Show all your work
- organize the problem in order.
- box your answer!!

① Limiting Distributions



$N \longrightarrow \infty$

the resulting distribution is called

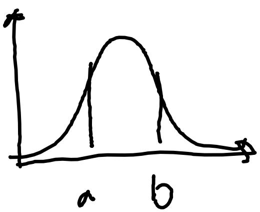
"limiting distribution"

$f(x)$: Probability Density Function (PDF)

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\bar{x} = \int_{-\infty}^{\infty} x\, f(x)\,dx$$

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x - \bar{x})^2 f(x)\,dx$$

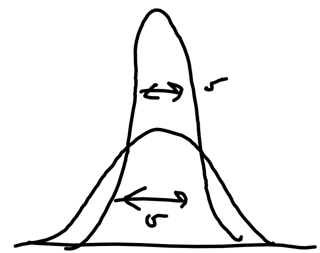


$$\int_a^b f(x)\,dx$$

(2) Normal Distribution

→ most important limiting Distribution

$$G_{X,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-X)^2/2\sigma^2}$$



X : center parameter

σ : width parameter

$$\text{Prob}(\text{with } t\sigma) = \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-z^2/2}\, dz}_{\text{error function}}$$

we use a table for it.

(3) Principle of Maximum Likelihood

$$\{x_1 \quad \cdots \quad x_N\}$$

We want to fit a normal distribution maximizes:

$$\text{Prob}_{\underbrace{\bar{X},\sigma}_{\text{parameters}}}(x_1, \ldots, x_N) \propto \frac{1}{\sigma^N} e^{-\sum (x_i - \bar{X})^2 / 2\sigma^2}$$

minimizing $(x_i - \bar{X})^2 / 2\sigma^2$

4. Rejection of Data

$$t_{sus} = \frac{|X_{sus} - \bar{X}|}{\sigma}$$

Prob(outside $t_{sus} \cdot \sigma$) $\rightarrow$ lookup on table.

$n = N \cdot \text{Prob}(\cdot) < 1/2 \Rightarrow$ outlier.

Chauvenet's criterion.

5. Weighted Average

Q: How to combine multiple measurments w/ different uncertainties?

$$X_{wav} = \frac{w_A X_A + w_B X_B}{w_A + w_B}$$

$$w_A = \frac{1}{\sigma_A^2}$$

$$w_B = \frac{1}{\sigma_B^2}$$

$$\Rightarrow X_{wav} = \frac{\sum_i w_i X_i}{\sum w_i}$$

$$w_i = \frac{1}{\sigma_i^2}$$

(6.) Least Squares $\quad y = A + Bx$

$$\chi^2 = \sum \frac{(y_i - A - Bx_i)^2}{\sigma_y^2}$$

$$\left\{ \begin{aligned} \frac{\partial \chi^2}{\partial A} &= -\frac{2}{\sigma_y^2} \sum (y_i - A - Bx_i) = 0 \\ \frac{\partial \chi^2}{\partial B} &= -\frac{2}{\sigma_y^2} \sum x_i (y_i - A - Bx_i) = 0 \end{aligned} \right.$$

$$\left\{ \begin{aligned} A \cdot N + B \sum x_i &= \sum y_i \\ A \cdot \sum x_i + B \sum x_i &= \sum x_i y_i \end{aligned} \right\} \text{ Normal Eq.}$$

$$A = \frac{\sum x^2 \sum y - \sum x \sum xy}{\Delta}$$

$$B = \frac{N \sum xy - \sum x \sum y}{\Delta}$$

$$\Delta = N \sum x^2 - \left(\sum x\right)^2$$

$$\sigma_y = \sqrt{\frac{1}{N-2} \sum_{i} (y_i - A - Bx_i)^2}$$

$$\sigma_A = \sigma_y \sqrt{\frac{\sum x^2}{\Delta}}$$

$$\sigma_B = \sigma_y \sqrt{\frac{N}{\Delta}}$$

Matrix Version:

$y = A + Bx$

$$Y = X\beta$$

β: parameter vector

X: data matrix

Y: true vector

$$\begin{Bmatrix} y_i \\ \vdots \\ \vdots \\ y_N \end{Bmatrix} = \begin{bmatrix} x_i & 1 \\ \vdots & \vdots \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix} \begin{bmatrix} B \\ A \end{bmatrix}$$

$N \times 1$ $\quad$ $N \times 2$ $\quad$ 2×1

$$\beta^{+} = \Big[\overbrace{(X^T X)^{-1} X^T}^{\text{pseudo-invers (pinv).}} \Big] Y$$

$$y = A\sin(x) + B$$

$$\begin{Bmatrix} y_i \\ \vdots \\ y_N \end{Bmatrix} = \begin{bmatrix} \sin(x_i) & 1 \\ \vdots & \vdots \\ \sin(x_N) & \vdots \end{bmatrix} \begin{Bmatrix} A \\ B \end{Bmatrix}$$

$$y = A_0 + A_1 x + A_2 x^2 + A_3 x^3$$

$$\begin{Bmatrix} y_1 \\ \vdots \\ y_N \end{Bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ \vdots & & & \\ 1 & x_N & x_N^2 & x_N^3 \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \\ A_2 \\ A_3 \end{bmatrix}$$

$$\begin{bmatrix} A_0^* \\ A_1^* \\ A_2^* \\ A_3^* \end{bmatrix} = \left[\left(X^T X\right)^{-1} X^T \right] Y$$