CV 510^{0}_{1}

Modeling, Uncertainty, and Data for Engineers

(July – Nov 2025)

Dr. Prakash S Badal

Flow

- Announcement
- Summary of Part-B

Announcements

- Today's lab on sampling and reliability
- Part 3 starts from tomorrow (9^{th} Oct)
  - signal processing
  - time series
  - machine learning
- Make-up exam?

Why should you care?

Why should you care about Risk & Reliability?

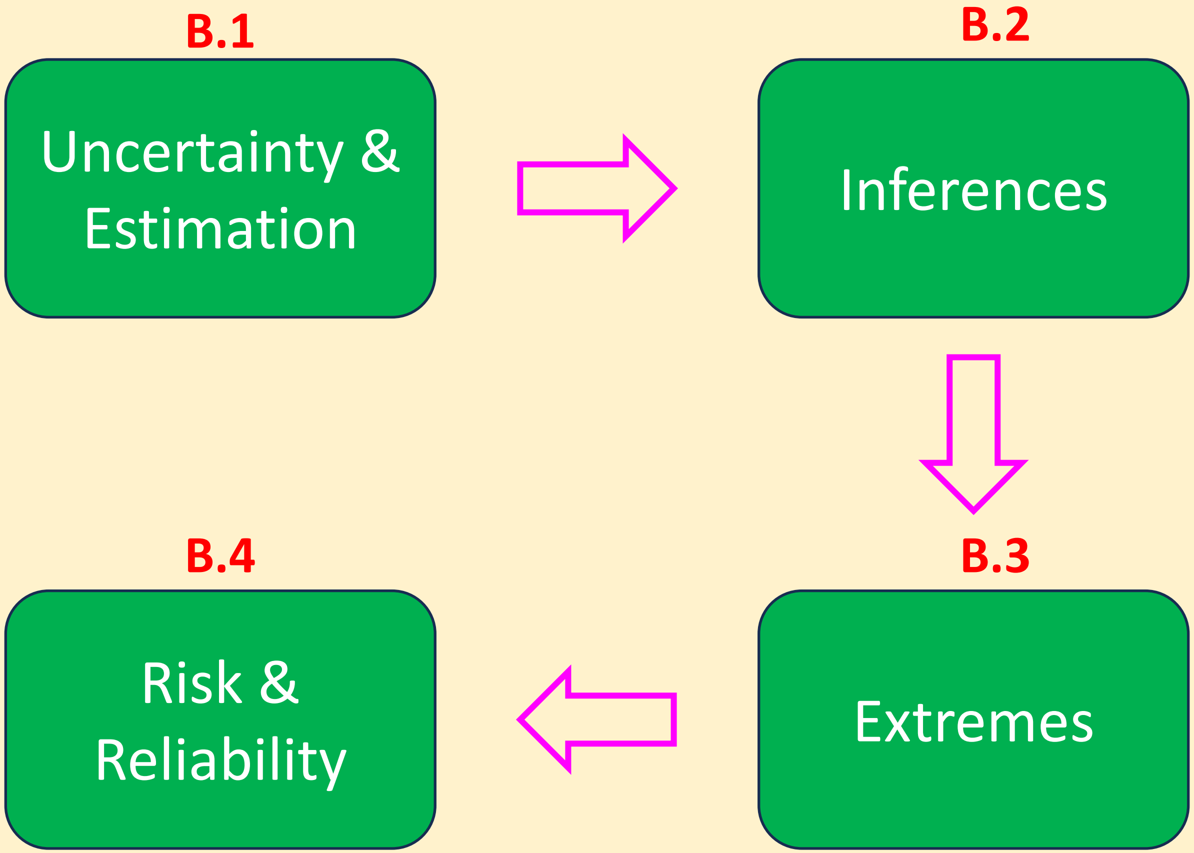
- Space Shuttle *Challenger* disaster (1986) https://www.youtube.com/watch?v=yibNEcn-4yQ

Catch-all approach:

Build for every "what if" scenario: create backup for every imaginable scenario, even wildly unlikely ones

If everything is important, then nothing is!

Module Overview



A crash course in probability, probabilistic models, and probabilistic methods.

Summary
of
Uncertainty & Estimation

Probability basics and RVs

- S: sample space; E: events in a random experiment
- Probability axioms are

$$\Pr(S) = 1$$

$$0 \leq \Pr(E) \leq 1$$

For mutually exclusive E_1 and E_2, $\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2)$

- Conditional probability: $\Pr(E_1|E_2) = \dfrac{\Pr(E_1 E_2)}{\Pr(E_2)}$
- Multiplication rule: $\Pr(E_1 E_2) = \Pr(E_1|E_2) \cdot \Pr(E_2)$
- Statistically independence (SI):

"Conditional probability of one event given the other has occurred" is identical to "its marginal probability." $\Pr(E_1|E_2) = \Pr(E_1)$

For SI events, $\Pr(E_1 E_2) = \Pr(E_1) \cdot \Pr(E_2)$

Settlers of Catan

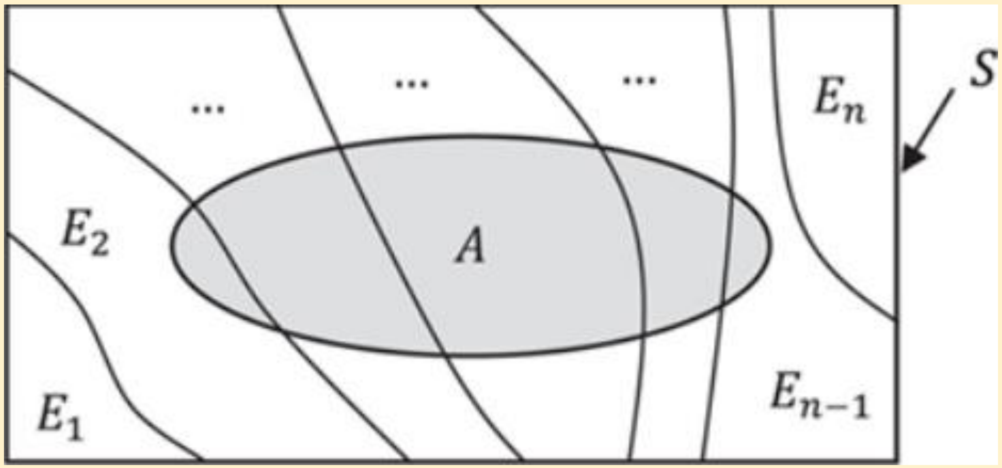
[Wikimedia commons]

What is the probability that you get at least one 8 in a single round consisting of four players?

colonist.io with permission

Probability rules: total probability rule

- If $E_1, E_2, \ldots, E_n$ are n MECE events,



$\Pr(A) =$

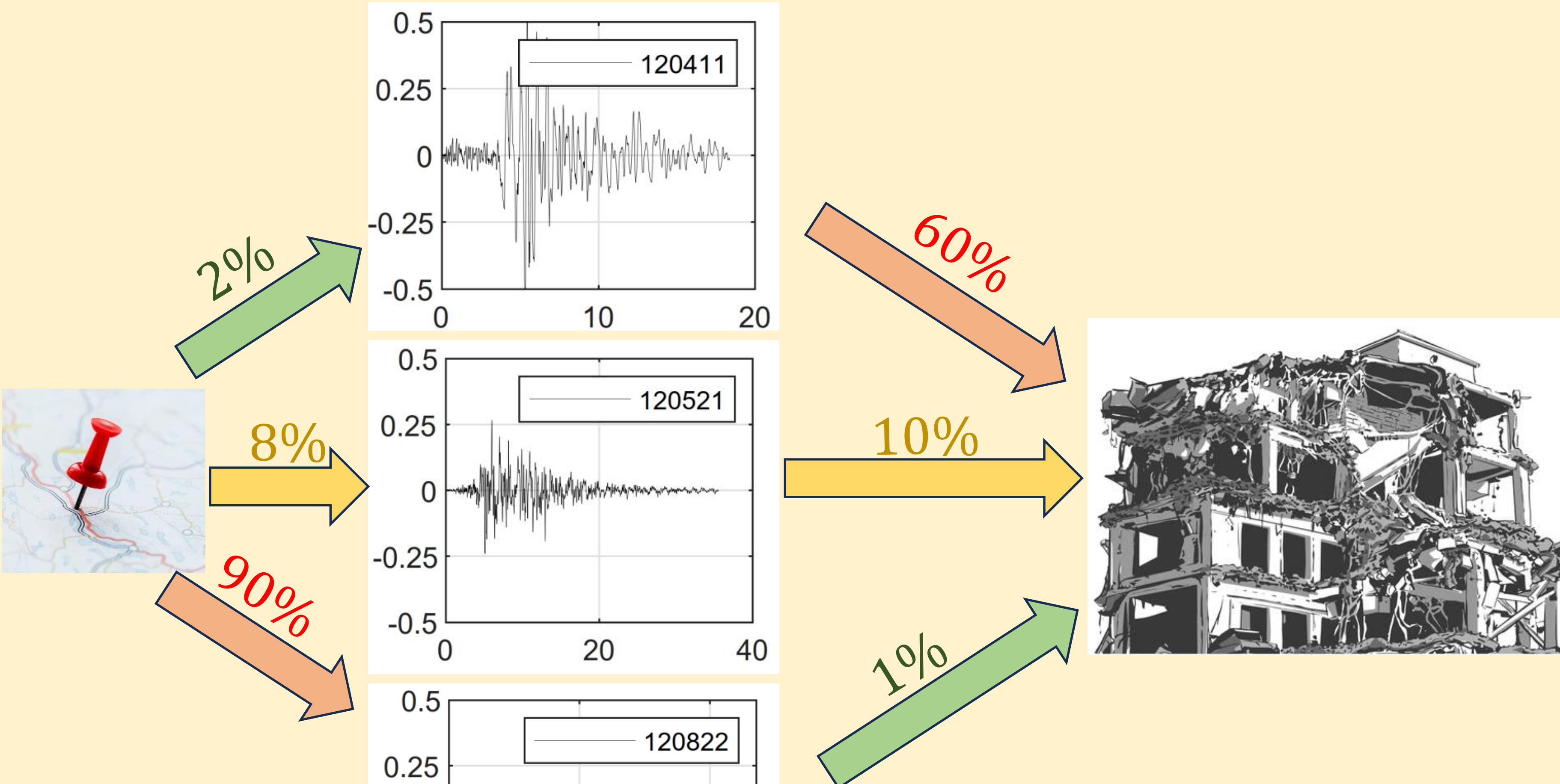
$$\Pr(AE_1) + \Pr(AE_2) + \cdots + \Pr(AE_n)$$

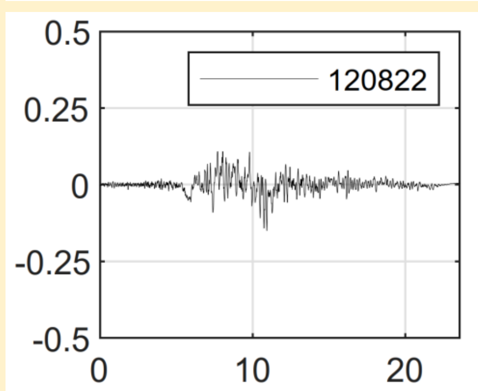
$$= \sum_{i=1}^{n} \Pr(AE_i)$$

$$= \sum_{i=1}^{n} \Pr(A|E_i)\Pr(E_i)$$

- Breaking down of calculation of event A into computing the conditional probabilities $P(A|E_i)$
- Conditionals usually easier to compute
- Clever selection of events E_i

Total Probability Rule





Denote earthquake intensity by W, M, and S; and damage by D.

$\Pr(W) = 0.90 \quad \Pr(M) = 0.08 \quad \Pr(S) = 0.02$

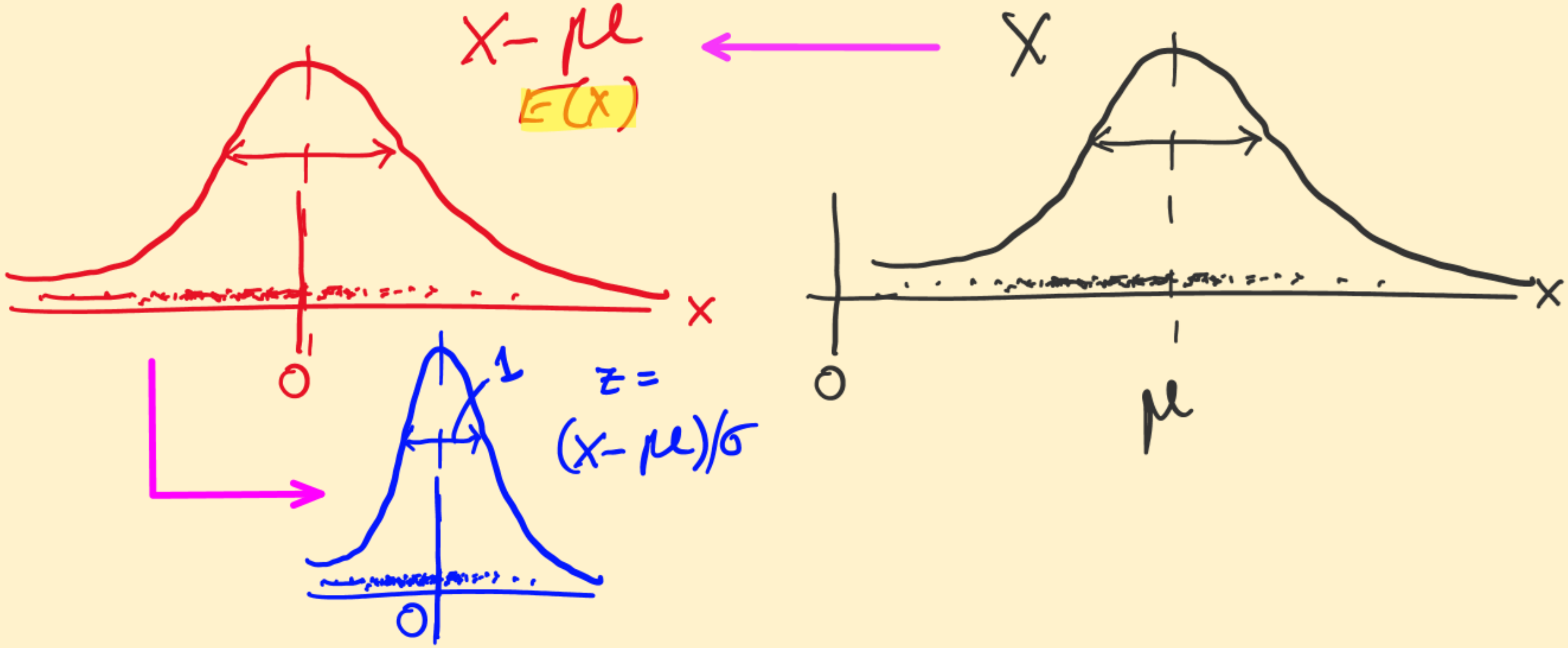
$\Pr(D|W) = 0.01 \quad \Pr(D|M) = 0.10 \quad \Pr(D|S) = 0.60$

$$\Pr(D) = \Pr(D \cap W) + \Pr(D \cap M) + \Pr(D \cap S)$$
$$= \Pr(D|W)\Pr(W) + \Pr(D|M)P(M) + \Pr(D|S)P(S)$$
$$= 0.009 + 0.008 + 0.012 = 0.029$$

Random variables, distributions, uncertainty propagation

Random variables: $\mathrm{E}[\cdot]$, $\mathrm{Var}[\cdot]$

Black dots → red dots → blue dots



dispersion ≡ Variance = $E[(X-\mu_X)^2]$

Covariance

- Covariance is the expected value of $(X_1 - \mu_1)(X_2 - \mu_2)$,

$\mathrm{Cov}[X_1, X_2] =$

$$\mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)]$$

$$\Rightarrow \mathrm{Cov}[X_1, X_2] = \mathrm{E}[X_1 X_2] - \mathrm{E}[X_1]\mathrm{E}[X_2]$$

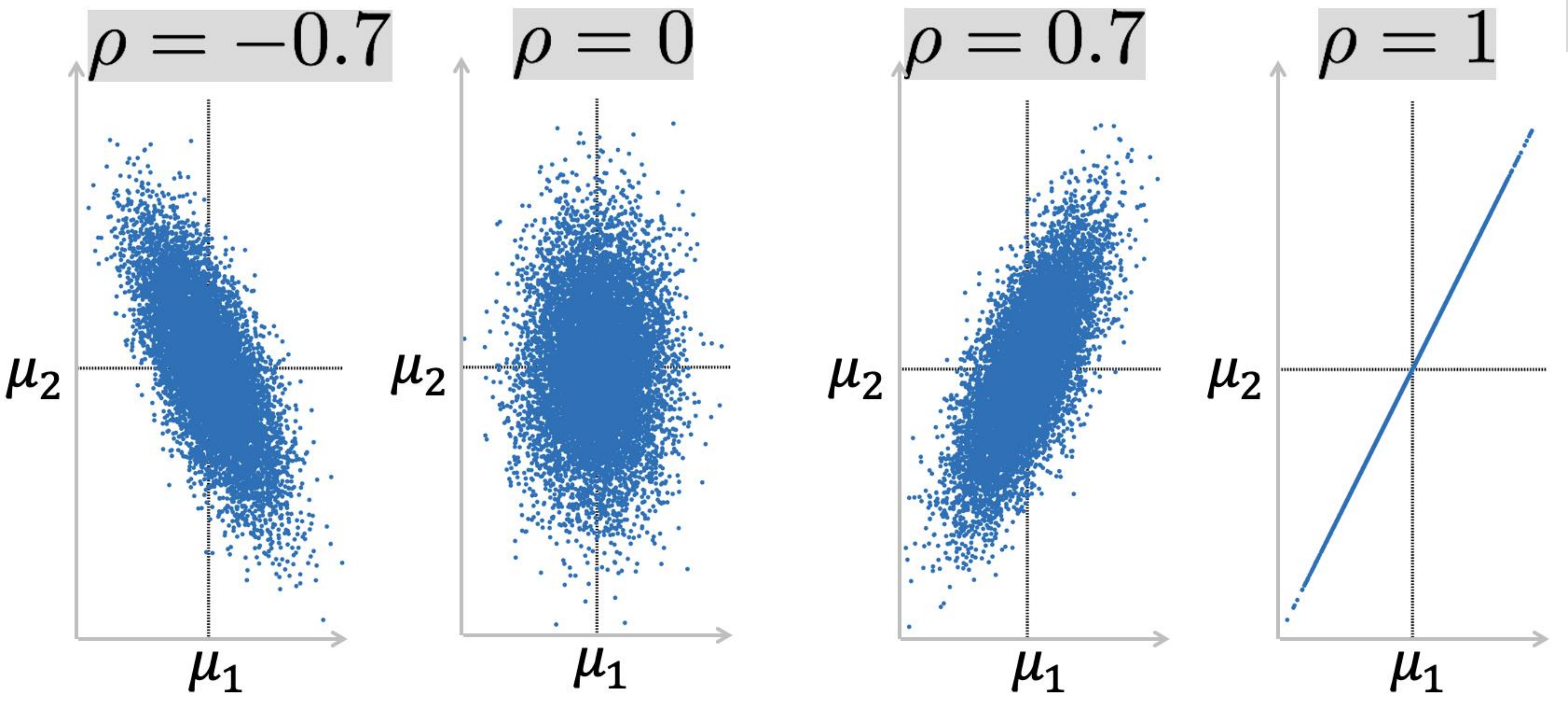
Correlation coefficient

$$\rho_{12} = \frac{\mathrm{Cov}[X_1, X_2]}{\sigma_1 \sigma_2}.$$

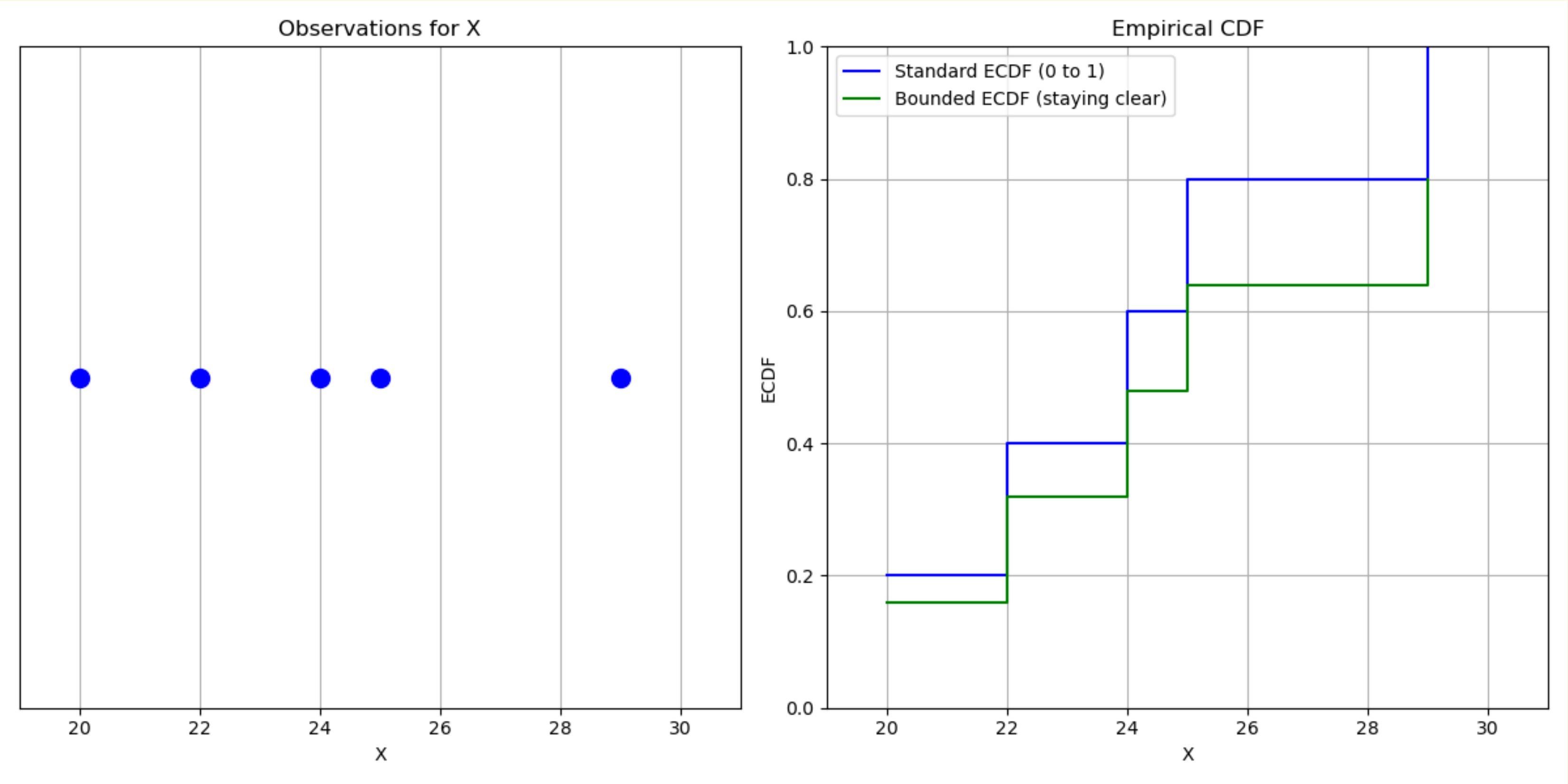
$$\mathrm{Cov}[X_1, X_2] = \rho_{12}\sigma_1\sigma_2$$

$$-1 \leq \rho_{12} \leq 1$$

Correlation



Empirical distribution



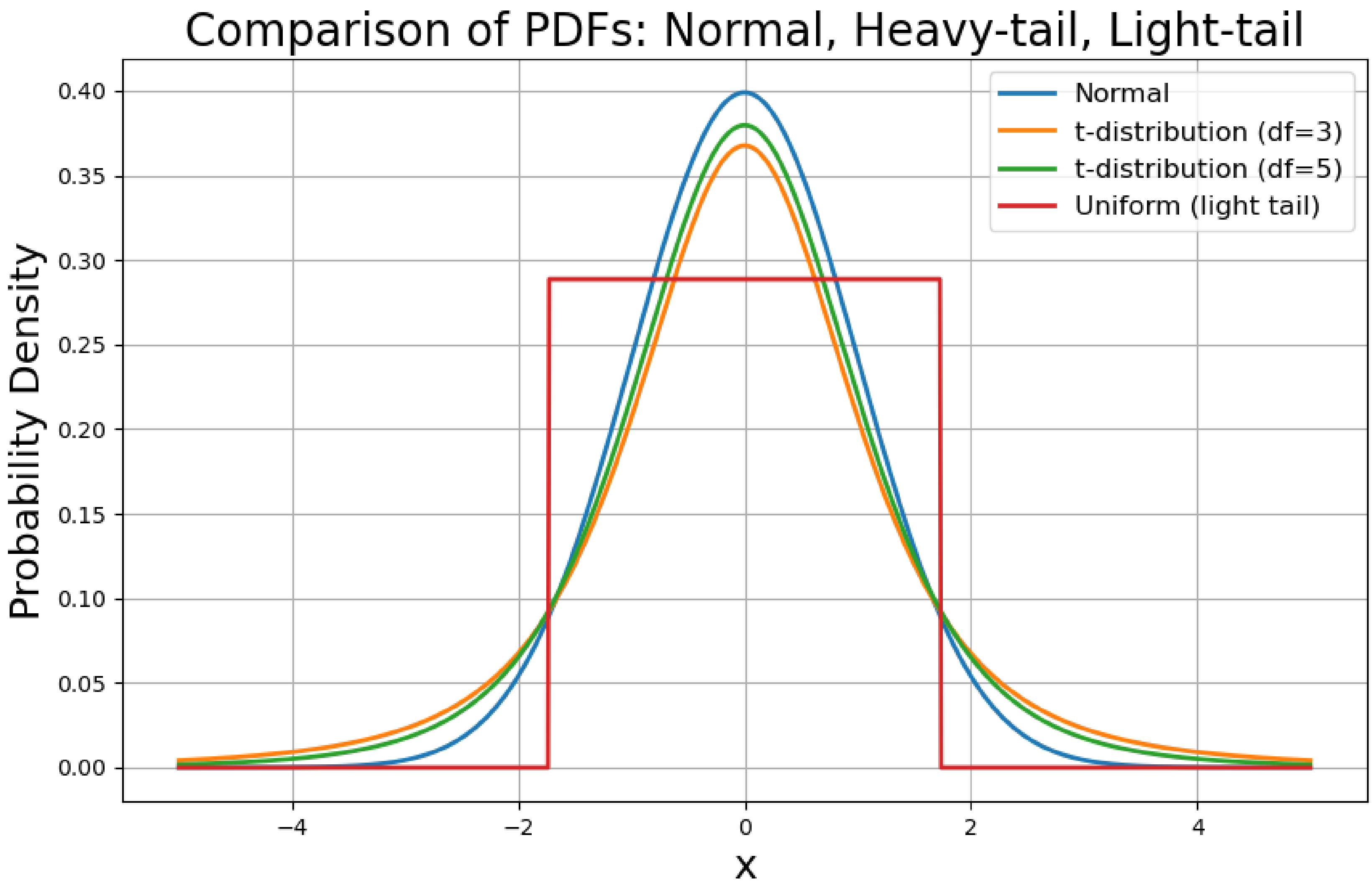
Standard ECDF: $F_n(x) = \frac{i}{n}$ goes from 0 to 1

Bounded ECDF: $F_{n,b}(x) = \frac{i}{n+1}$ stays clear of 0 and 1

useful in Q-Q/probability plotting, avoids $-\infty$ or $+\infty$.

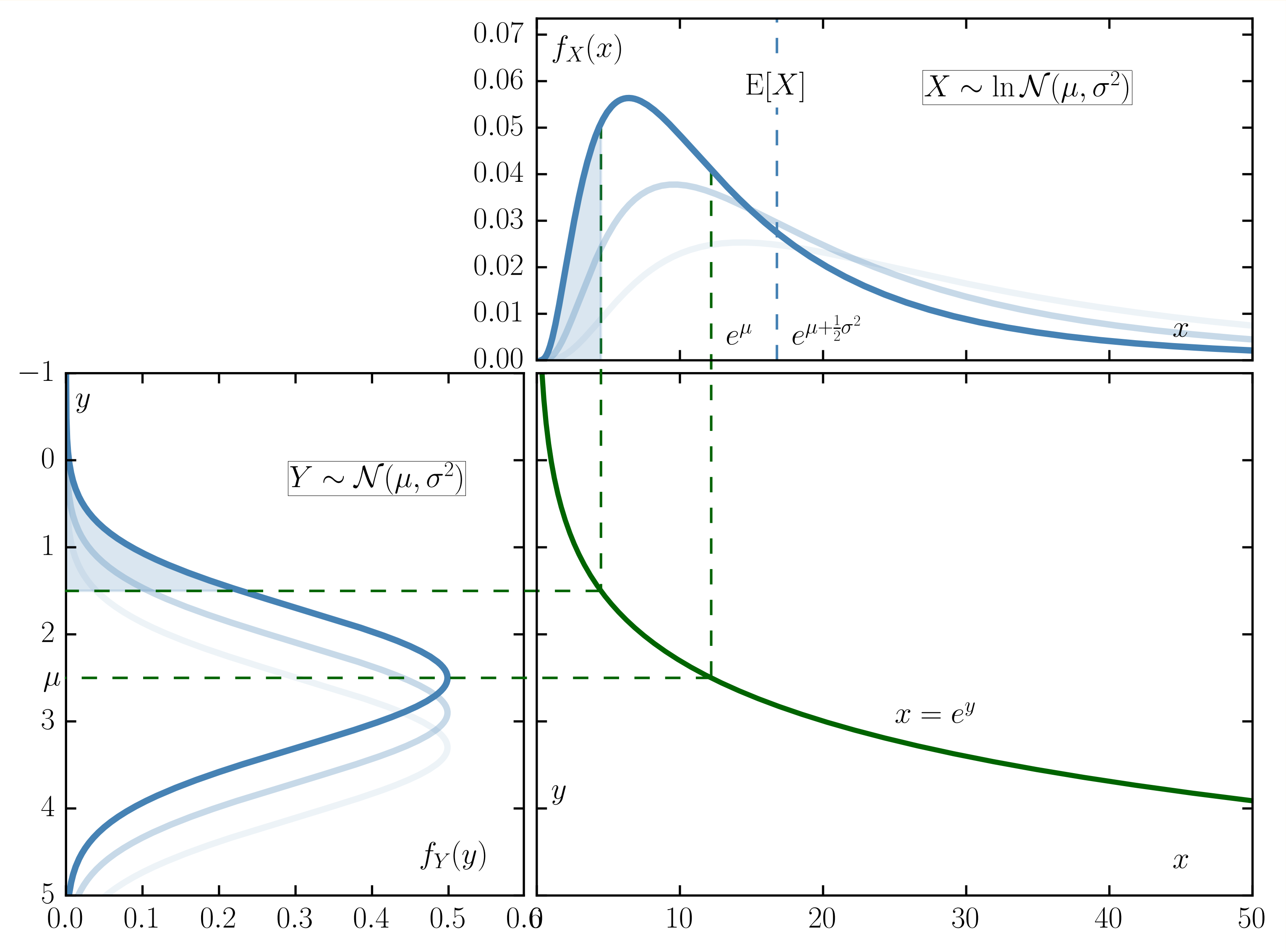
Core distributions

- *Tails*: thick/light



Lognormal

- $X \sim Lognormal \iff Y = \ln(X) \sim Normal$



Gumbel distributions

- When we are interested in

the smallest

or

the largest of a set of rv's,

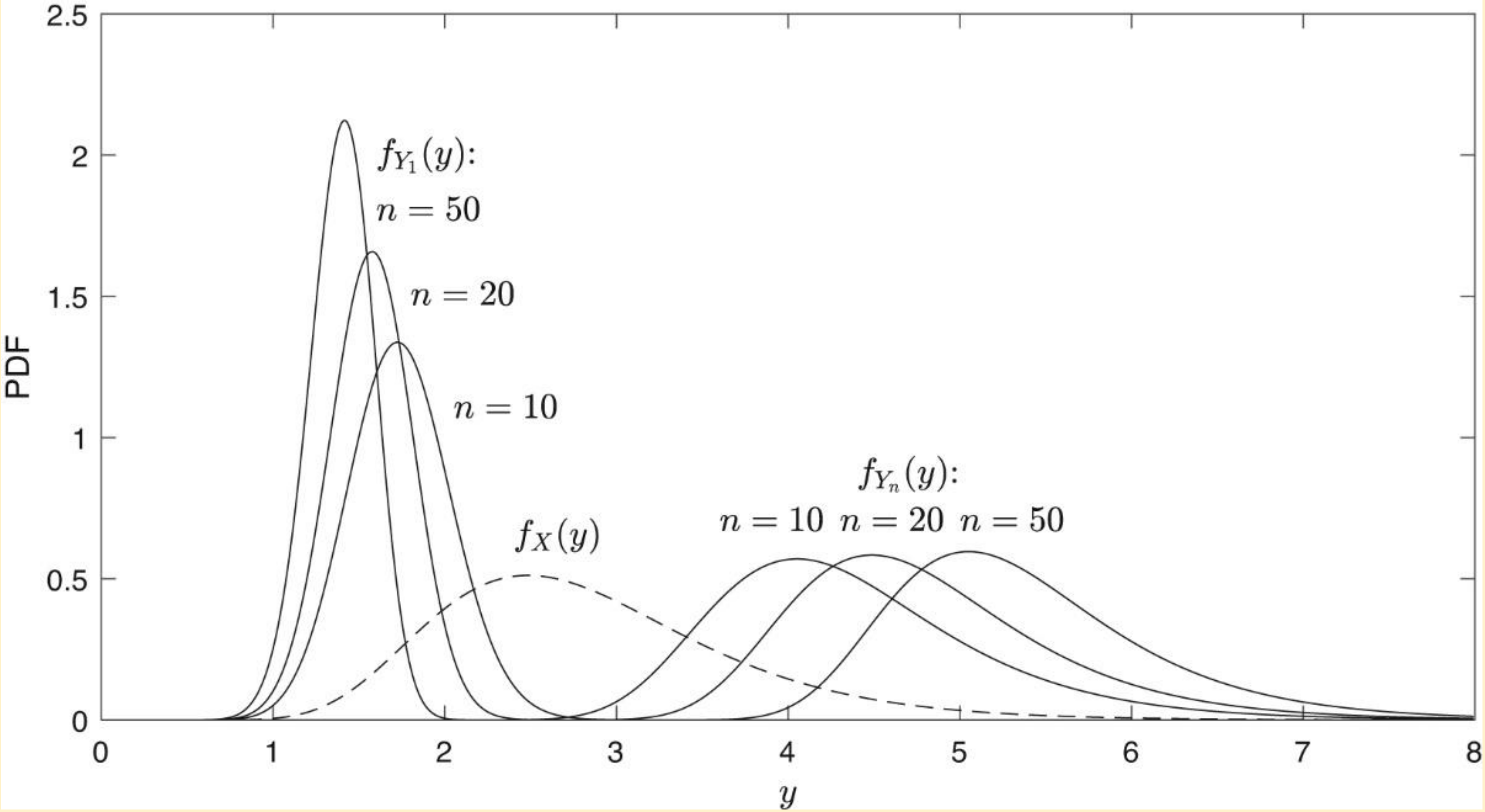
e.g., a chain of links: smallest strength.

Flood level under a bridge: highest flood level during its lifetime.

$$Y_1 = \min(X_1, X_2, \dots, X_n),$$

$$Y_n = \max(X_1, X_2, \dots, X_n).$$

Gumbel distributions



Error propagation

Revise

- Two structural engineers, Alice and Bob, independently measure the wind speed in m/s at the top of each tower.

Alice's reading $X \sim \mathcal{N}(\mu = 40, \sigma = 5)$

Bob's reading $Y \sim \mathcal{N}(\mu = 42, \sigma = 6)$

Their readings are correlated with $\rho = 0.8$

1. What's the probability that Alice's reading exceeds 50?
2. What's the covariance of X and Y?
3. If Alice and Bob average their measurements, $W = (X + Y)/2$, what are $\mathrm{E}[W]$ and $\mathrm{Var}[W]$?
4. If design wind speed is 55 m/s, what's the prob. that the average exceeds 55 m/s?

Mean and Variance propagation laws

If $Y = a_1 X_1 + a_2 X_2 + c$, with a_i and c deterministic const.

$\mathrm{E}[Y] = a_1\mu_1 + a_2\mu_2 + c$

$\mathrm{Var}[Y] = \quad \mathrm{E}[(Y - \mu_Y)^2]$

$= \mathrm{E}[\{(a_1X_1 + a_2X_2 + c) - (a_1\mu_1 + a_2\mu_2 + c\ )\}^2]$

$= \mathrm{E}[\{(a_1X_1 - a_1\mu_1) + (a_2X_2 - a_2\mu_2)\}^2]$

$= \mathrm{E}[(a_1X_1 - a_1\mu_1)^2 + (a_2X_2 - a_2\mu_2)^2 + 2(a_1X_1 - a_1\mu_1)(a_2X_2 - a_2\mu_2)]$

$= a_1^2\mathrm{E}[(X_1 - \mu_1)^2] + a_2^2\mathrm{E}[(X_2 - \mu_2)^2] + 2a_1a_2E[2(X_1 - \mu_1)(X_2 - \mu_2)]$

$= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 2a_1a_2\mathrm{Cov}[X_1, X_2]$

Mean and Variance propagation laws

- If $Y = g(X)$ is a nonlinear function of X. Find $\mathrm{E}[Y]$ & $\mathrm{Var}(Y)$.

$\mathrm{E}[Y] = \mathrm{E}[g(X)]$

Taylor series expansion:

$$g(X) = g(\mu_X) + \left(\frac{\partial g}{\partial x}\right)_{\mu_X} (X - \mu_X) + \frac{1}{2!}\left(\frac{\partial^2 g}{\partial x^2}\right)_{\mu_X} (X - \mu_X)^2 + \mathrm{H.O.T.}$$

$$\mathrm{E}[Y] \cong \mathrm{E}\left(g(\mu_X) + \left(\frac{\partial g}{\partial x}\right)_{\mu_X} (X - \mu_X) + \frac{1}{2!}\left(\frac{\partial^2 g}{\partial x^2}\right)_{\mu_X} (X - \mu_X)^2\right)$$

$$= g(\mu_X) + 0 + \frac{1}{2!}\left(\frac{\partial^2 g}{\partial x^2}\right)_{\mu_X} \mathrm{E}[(X - \mu_X)^2]$$

$\mathrm{E}[Y] \cong g(\mu_X)$ First-order mean approximation

$\mathrm{E}[Y] \cong g(\mu_X) + \frac{1}{2}\left(\frac{\partial^2 g}{\partial x^2}\right)_{\mu_X} \sigma_X^2$ Second-order mean approximation

Mean and Variance propagation laws

- If $Y = g(X)$ is a nonlinear function of X. Find $\mathrm{E}[Y]$ & $\mathrm{Var}(Y)$.

Taylor series expansion:

$$g(X) = g(\mu_X) + \left(\frac{\partial g}{\partial x}\right)_{\mu_X}(X - \mu_X) + \frac{1}{2!}\left(\frac{\partial^2 g}{\partial x^2}\right)_{\mu_X}(X - \mu_X)^2 + \mathrm{H.O.T.}$$

$$\mathrm{Var}[Y] = \mathrm{E}[(Y - \mu_Y)^2] \cong \left(\left(\frac{\partial g}{\partial x}\right)_{\mu_X}\right)^2 \sigma_X^2$$

First-order var. approx.

Regression & estimation

Elements of models

0. Eyeball the data.

Scatter plot, histogram, change scales

1. Estimation

Goal: Obtain parameter estimates ($\hat{\beta}$)

Concepts: least squares, maximum likelihood, fitting the model

2. Inference

Goal: Model comparison; uncertainty in parameter ($\hat{\beta}$)

Concepts: Conf. interval for $\hat{\beta}$, hyp. testing, std. error, p-values

3. Prediction

Goal: Forecast new outcomes (x is now a predictor)

Concepts: CI for $\hat{y}$ (prediction error), mean-squared error (MSE)

4. Explanation

Goal: Interpret the fitted model, understand relationships

Concepts: feature importance, causality

5. Diagnosis

Goal: Assess model assumptions and validity

Concepts: error (constant Var.), unusual observations (outliers)

Elements of models: Estimation

0. Eyeball the data.

Scatter, histogram, change scales (log-log, semilogX, semilogY, e^X, so on).

1. Estimation

2. Inference

3. Prediction

4. Explanation

5. Diagnosis

Model: $Y = Ax + \epsilon$ $Y = x\beta + \epsilon$

~~Predicted~~/response: Y dependent var.

~~Predictor~~/feature: x ind./explanatory/covariate

Parameters: A or β

Estimating parameters aka "model building stage":

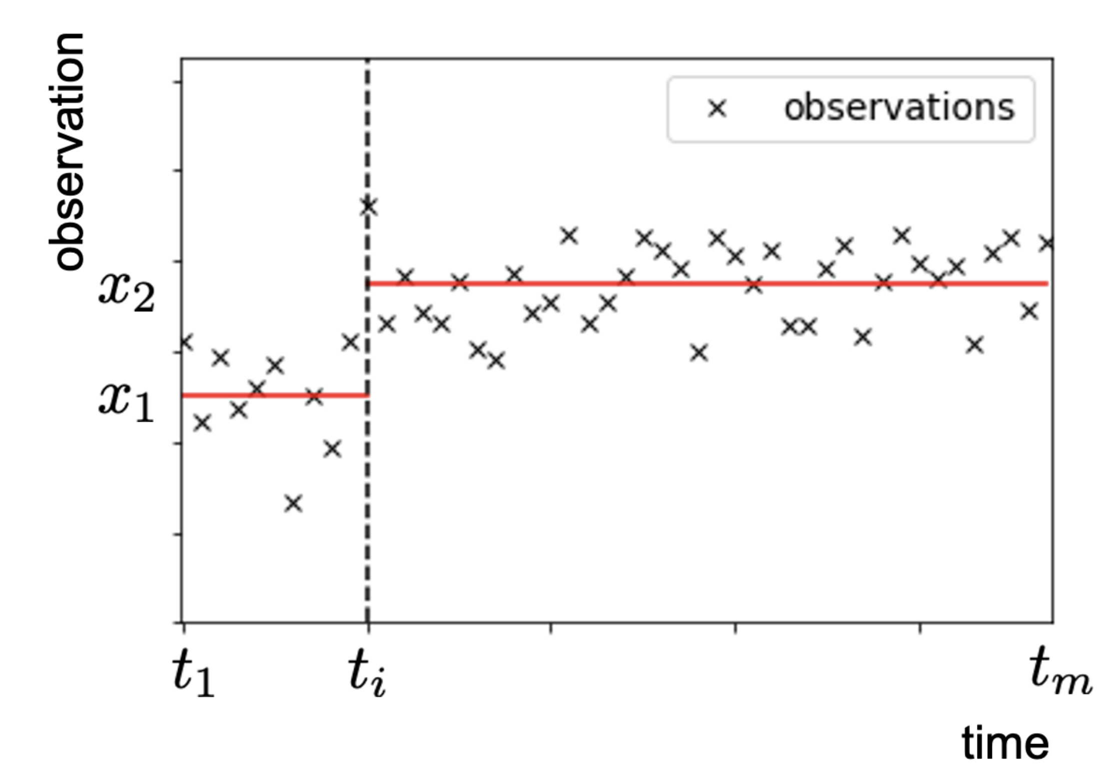
- Role of x is not to predict (y) as yet!
- It is to estimate A or β
- Better call x at this stage feature/covariate/explanatory var.

Goal of estimation: Obtain parameter estimates ($\widehat{A}$)

Linear models: example

- Is this a linear model

Write its linear functional relationship



$$E\left(\begin{bmatrix} Y_1 \\ \vdots \\ Y_{i-1} \\ Y_i \\ \vdots \\ Y_n \end{bmatrix}\right) = \underbrace{\begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{x}$$

A linear model in x

Least-square errors

A linear model with a single feature has two parameters:

Intercept slope

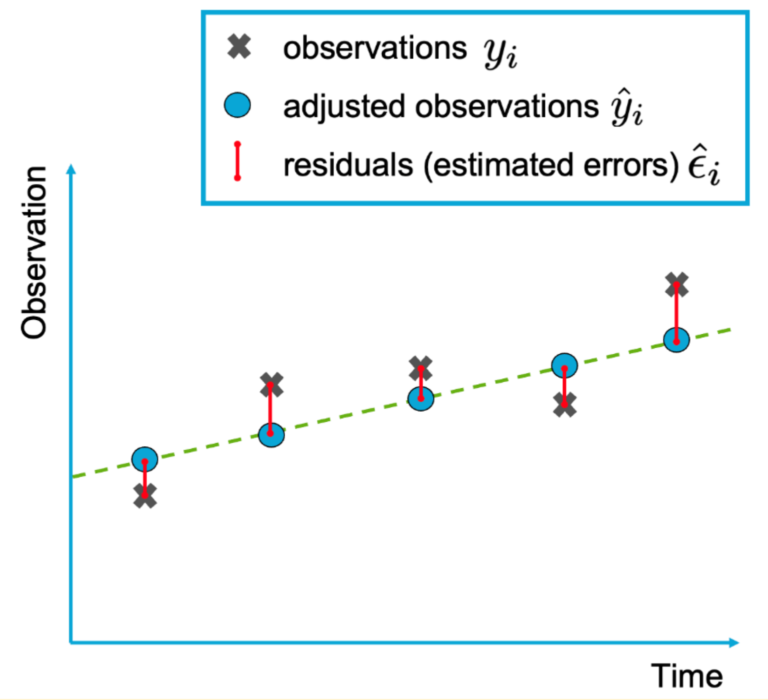
An example:

$y_1 = mx_1 + c$

…

$y_5 = mx_5 + c$

$$A = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}; x = \begin{bmatrix} m \\ c \end{bmatrix}$$



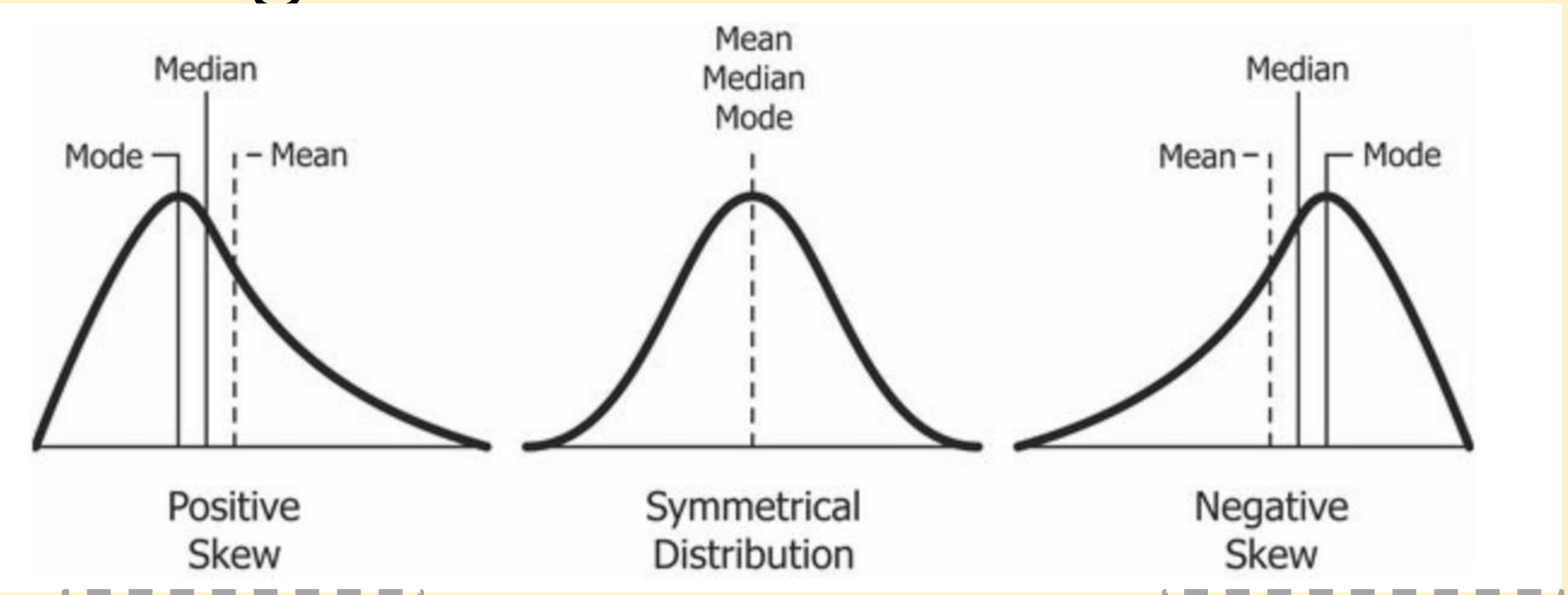
5 equations; 2 unknowns (m, c)

$$A^{\mathrm{T}}y = A^{\mathrm{T}}Ax \Rightarrow x = \left(A^{\mathrm{T}}A\right)^{-1}A^{\mathrm{T}}y$$

~~Magic:~~ What does the calculated m and c mean?

Least-square estimates of m, c.

Exploiting central moments



Right skew
(skewness > 0)

Left skew
(skewness < 0)

Central moments:

First:	$\mathrm{E}[(X-\mu)]$	zero
Second:	$\mathrm{E}[(X-\mu)^2]$	Variance
Third:	$\mathrm{E}[(X-\mu)^3]$	scaled Skewness (divide by σ^3 to get Skewness)
Fourth:	$\mathrm{E}[(X-\mu)^4]$	scaled Kurtosis (divide by σ^4 to get Kurtosis)

$$skewness = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{E[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}$$

Exploiting central moments

S. No.	Distribution	Skewness	Kurtosis (excess)	#param.
1	Normal $\mathcal{N}(\mu, \sigma)$	0	3 (0)	2
2	Uniform $[a, b]$	0	1.8 (-1.2)	2
3	Exponential $\mathrm{Exp}(\lambda)$	2	9 (+6)	1
4	Lognormal $\mathcal{LN}(\lambda, \zeta)$	Bad-looking fun of ζ	Bad-looking fun of ζ (above -3)	2
5	Gumbel type-1 (largest)	≈ 1.14	5.4 (+2.4)	2
6	Gumbel type-2 (smallest)	≈ -1.14	5.4 (+2.4)	2
7	t-distribution	0	$\frac{3(n-2)}{n-4}$; excess of $\frac{6}{n-4}$	2, #dof
8	Weibull	Param-dependent	Param-dependent	3
9	Beta	Param-dependent	Param-dependent	4 (2 shape, loc, scale)

Maximum Likelihood Estimation

Maximum likelihood estimation

Find $\mathcal{L}(\mu,\sigma|x_i) = f(x_i|\mu,\sigma)$ for each dart, x_i.

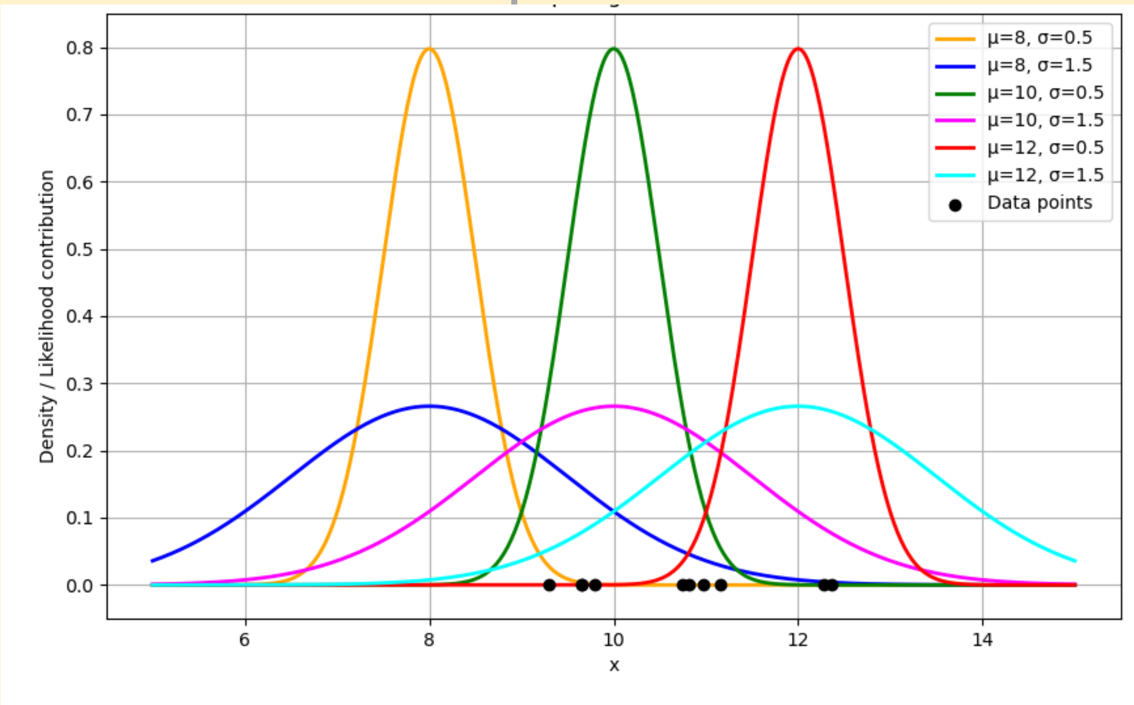
Maximize the product of likelihood,

$$\mathcal{L}(\mu,\sigma|\mathbf{x}) = \prod_{i}^{n} f(x_i|\mu,\sigma)$$

$$\hat{\mu},\hat{\sigma} = \arg\max_{\mu,\sigma} \mathcal{L}(\mu,\sigma|\mathbf{x})$$

$$\hat{\mu},\hat{\sigma} = \arg\max_{\mu,\sigma} \ln[\mathcal{L}(\mu,\sigma|\mathbf{x})]$$

Excel Demo!



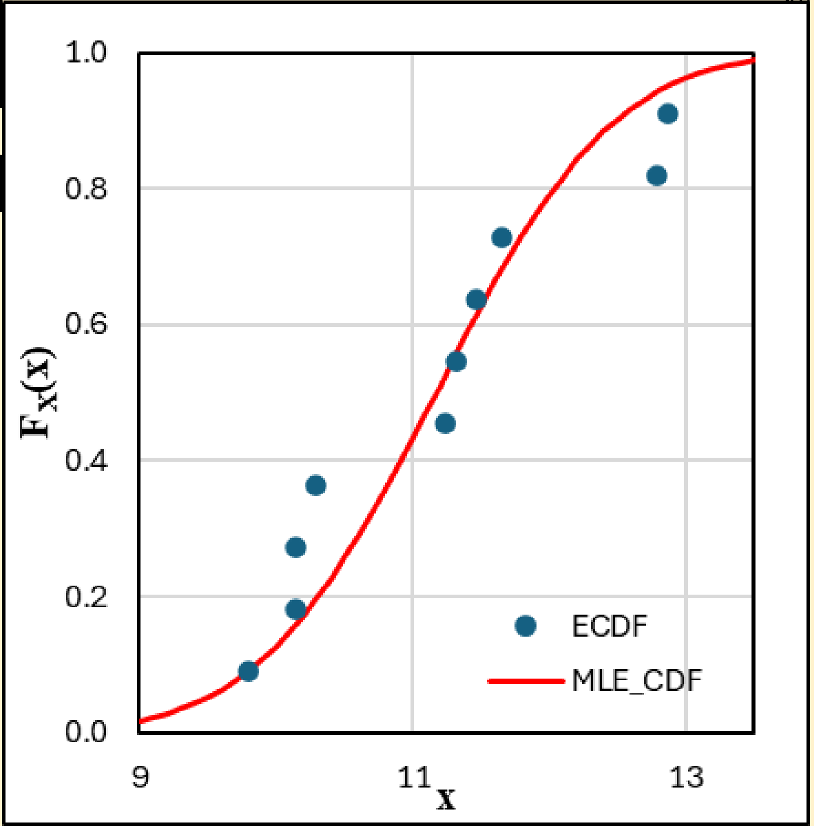
color	oran.	blue	green	mag.	red	turq.
μ_guess	**8**	**8**	**10**	**10**	**12**	**12**
σ_guess	**0.5**	**1.5**	**0.5**	**1.5**	**0.5**	**1.5**
x_i	Likelihood					
10.7	0.00	0.05	0.26	0.24	0.03	0.19
9.8	0.00	0.13	0.73	0.26	0.00	0.09
11.0	0.00	0.04	0.12	0.22	0.10	0.21
12.3	0.00	0.00	0.00	0.08	0.68	0.26
9.6	0.00	0.15	0.62	0.26	0.00	0.08
9.6	0.00	0.15	0.62	0.26	0.00	0.08
12.4	0.00	0.00	0.00	0.08	0.61	0.26
11.2	0.00	0.03	0.06	0.20	0.19	0.23
9.3	0.03	0.18	0.30	0.24	0.00	0.05
10.8	0.00	0.05	0.21	0.23	0.05	0.19
Sum-log-lik.	-166	-31	-32	-17	-59	-20

*Use "solver" to **maximize** Sum-log-lik. by changing μ, σ.*

μ_goal-seek	**10.67**
σ_goal-seek	**1.03**

Lik.	Log-lik.
0.39	-0.4
0.27	-0.6
0.37	-0.4
0.11	-0.9
0.24	-0.6
0.24	-0.6
0.10	-1.0
0.35	-0.5
0.16	-0.8
0.38	-0.4

Sum-log-lik. **-6.3**



Precision and Bias

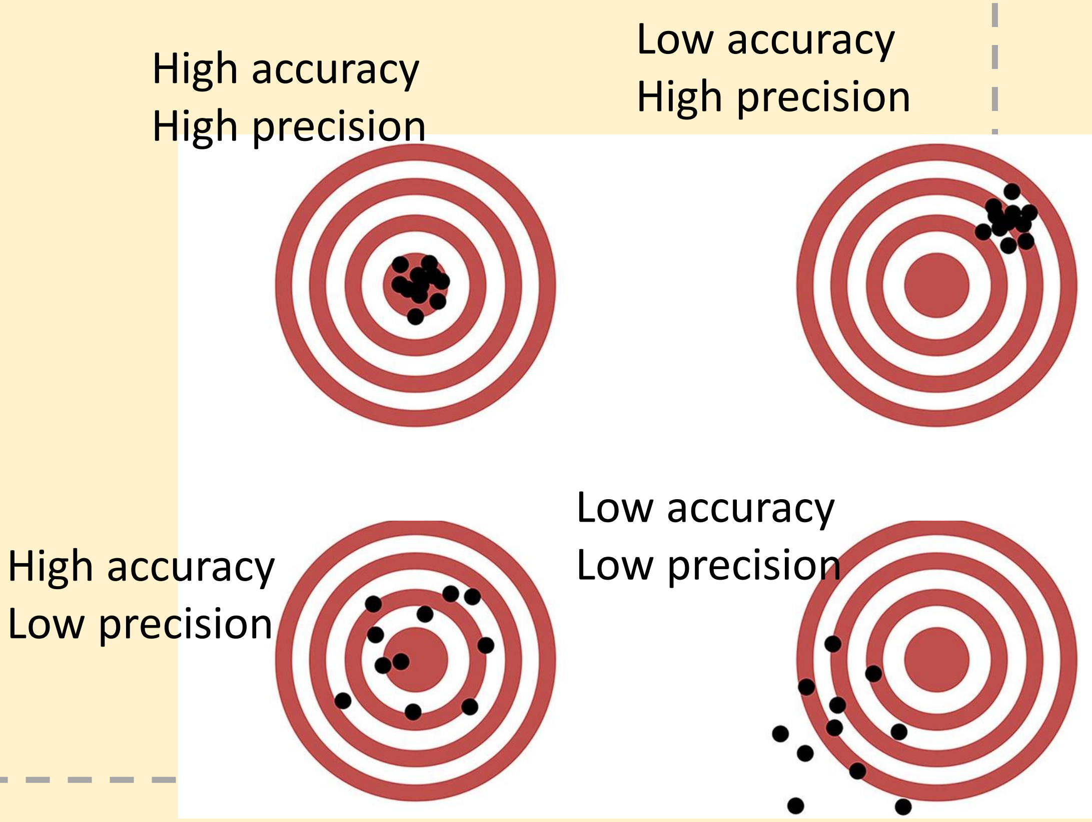
Bias and precision

- Accuracy/bias:

  How far off, on average, are your darts from the bullseye?

- Precision:

  How close are the darts to each other?



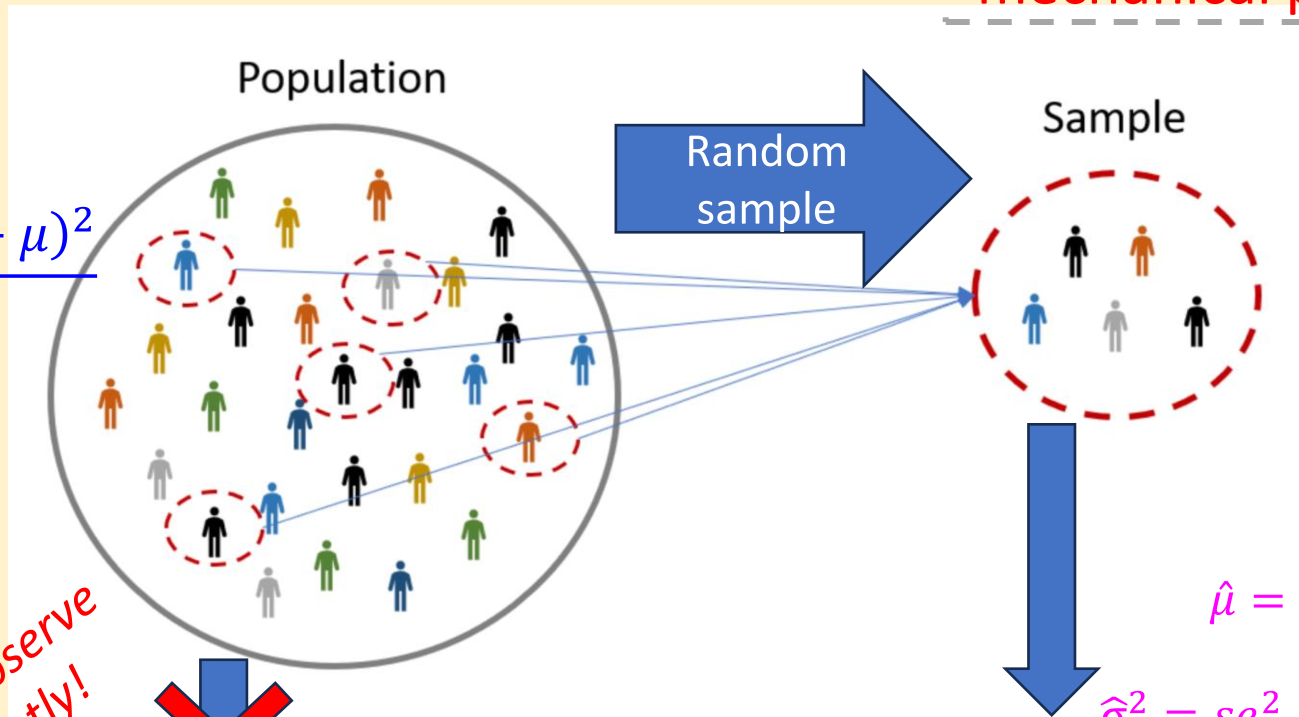
Confidence Interval Objective

Confidence interval

e.g., election polls, mechanical properties.

$$\mu = \frac{\sum_i^N x}{N}$$

$$\sigma^2 = \frac{\sum_i^N (x_i - \mu)^2}{N}$$



$$\hat{\mu} = \bar{X}_n = \frac{\sum_i^n x}{n}$$

$$\hat{\sigma}^2 = se^2 = \frac{\sum_i^n (x_i - \bar{X}_n)^2}{n-1}$$

Can't observe directly!

Population parameter: Population average

Inference

Sample statistic: Sample average

Use sample mean/proportion to estimate population mean/proportion

Confidence Interval Estimation

CLT and CI: (estimating mean)

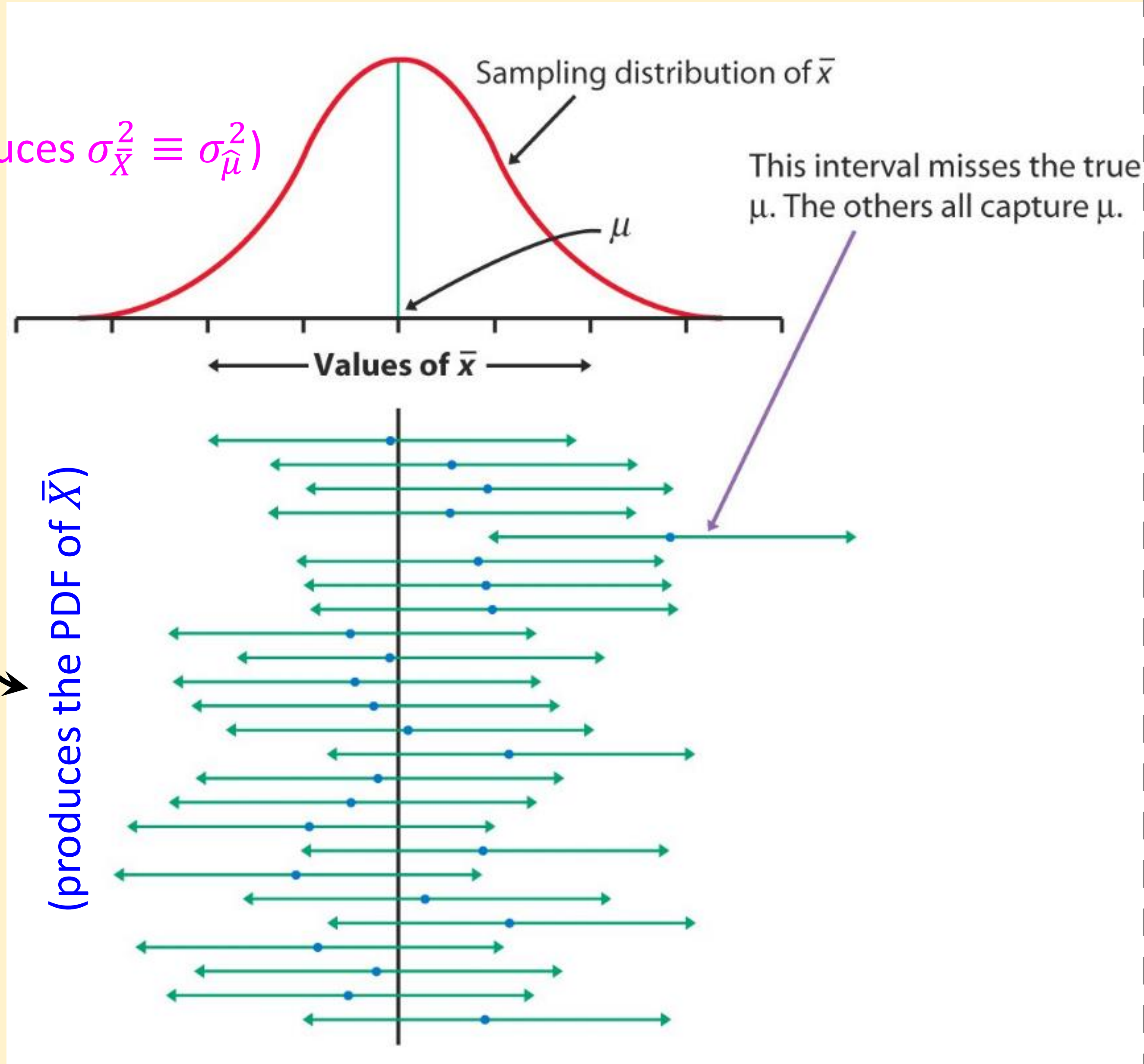
(reduces $\sigma_{\bar{X}}^2 \equiv \sigma_{\hat{\mu}}^2$)

Number of specimen
in a sample
(used for averaging)

vs.

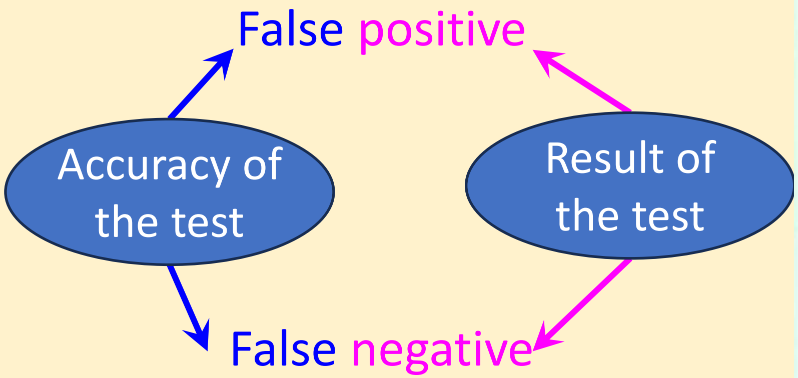
Number of reps

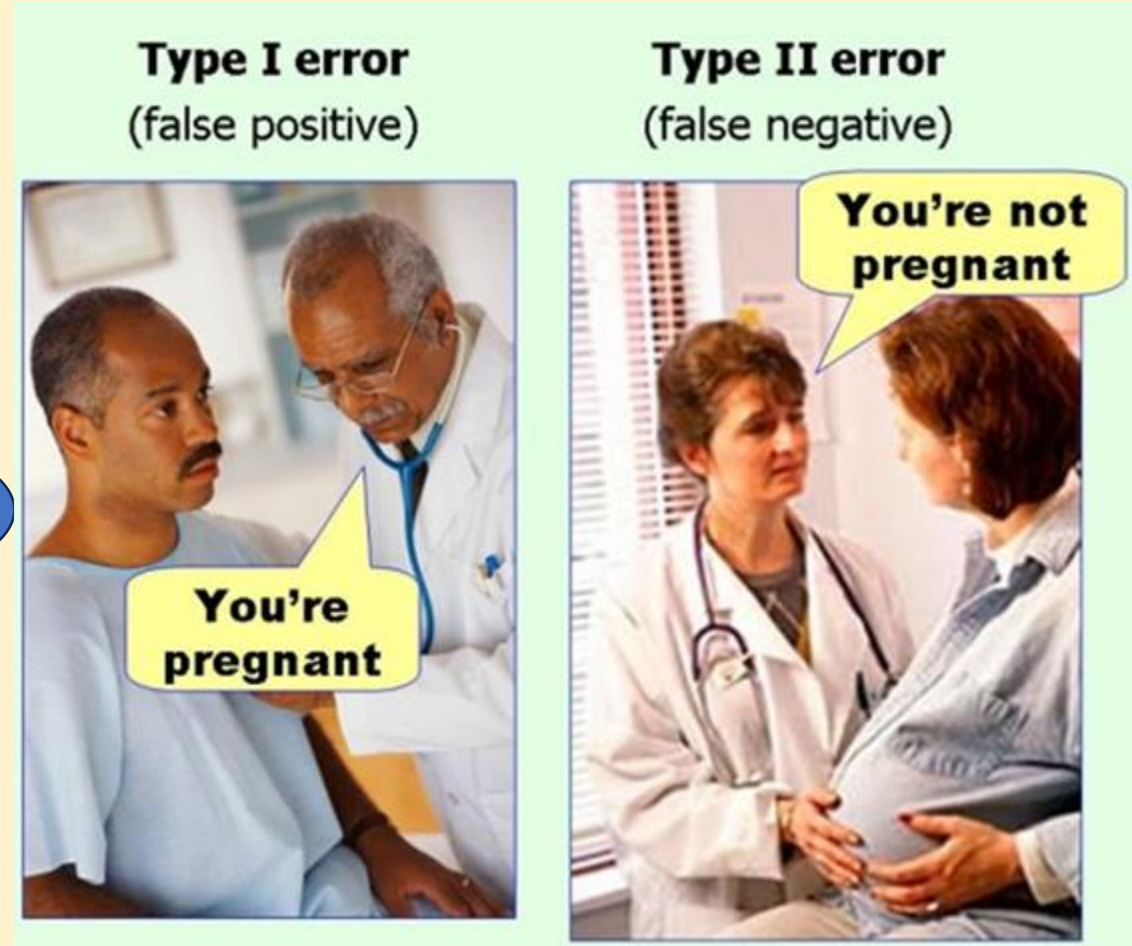
(produces the PDF of $\bar{X}$)



Hypothesis Testing

Spot the error*





People are generally not pregnant!

Default	You are not pregnant	H_0
Alternative	You are pregnant	H_1

Hypothesis testing (& justice system)

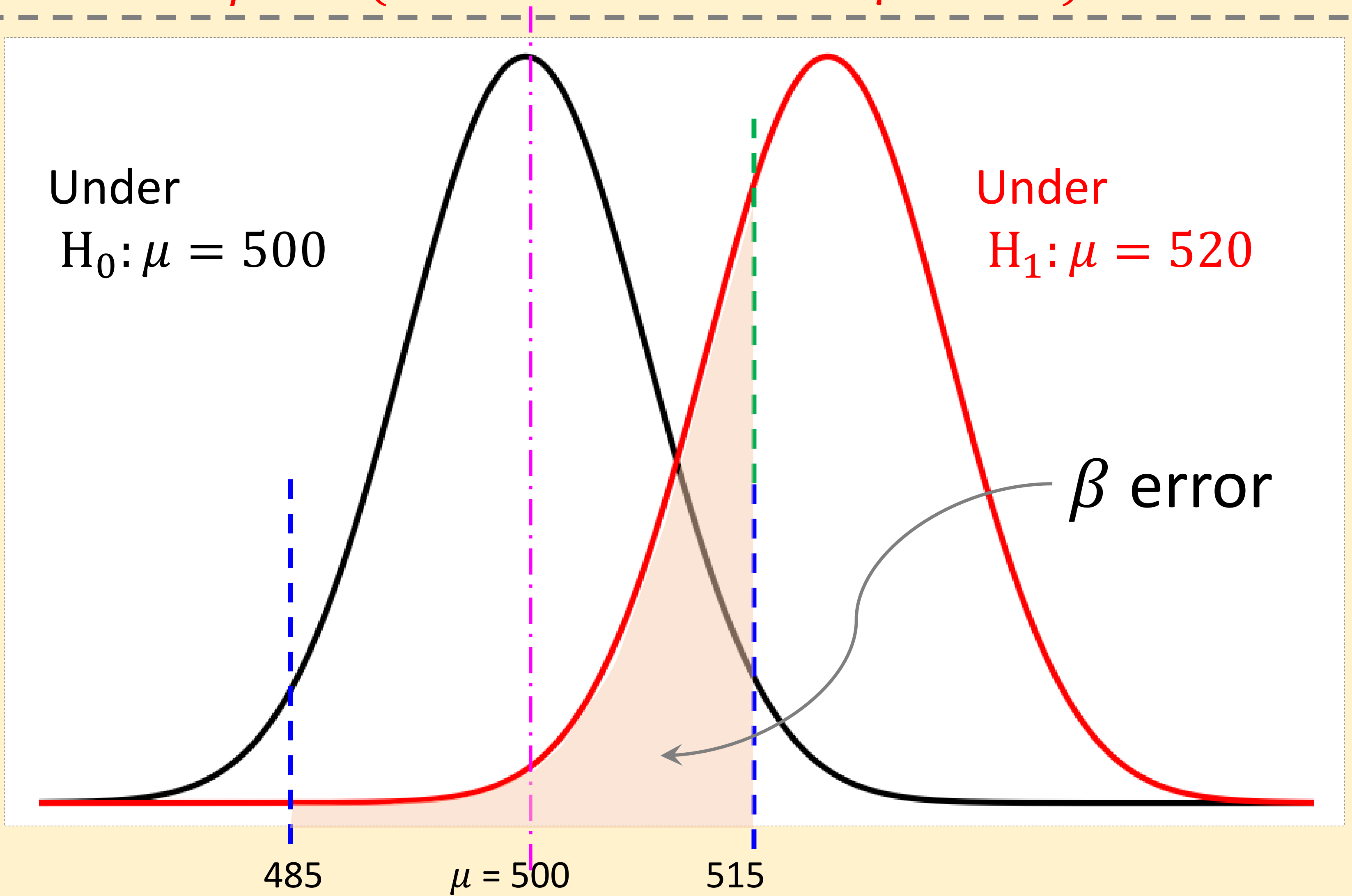
No numerical values in courts, but they share four common features:

1. **The alternative hypothesis:** This is why a *criminal is arrested*.
   - The police, of course, do not think that the criminal is innocent.
   - The researchers think that their treatment is effective. H_1 or H_A.
2. **The null hypothesis:** The *presumption of innocence*.
   - The suspect or treatment didn't do anything. H_0 is the logical opposite of H_1.
3. **A standard of justice:** *A reasonable doubt*. A test score!
   - No possibility of absolute proof. So, a standard has to be set.
   - Reject the null hypothesis beyond a reasonable doubt.
4. **A data sample:** Evaluation of *partial information.*
   - Eye-witnesses/fingerprints/DNA analysis/experimental/numerical data of treatment.
   - Getting the "whole truth and nothing but the truth" is often impossible.

Type II error

Type II error will be committed if $\bar{x} \in [485,515]$ when $\mu = 520$

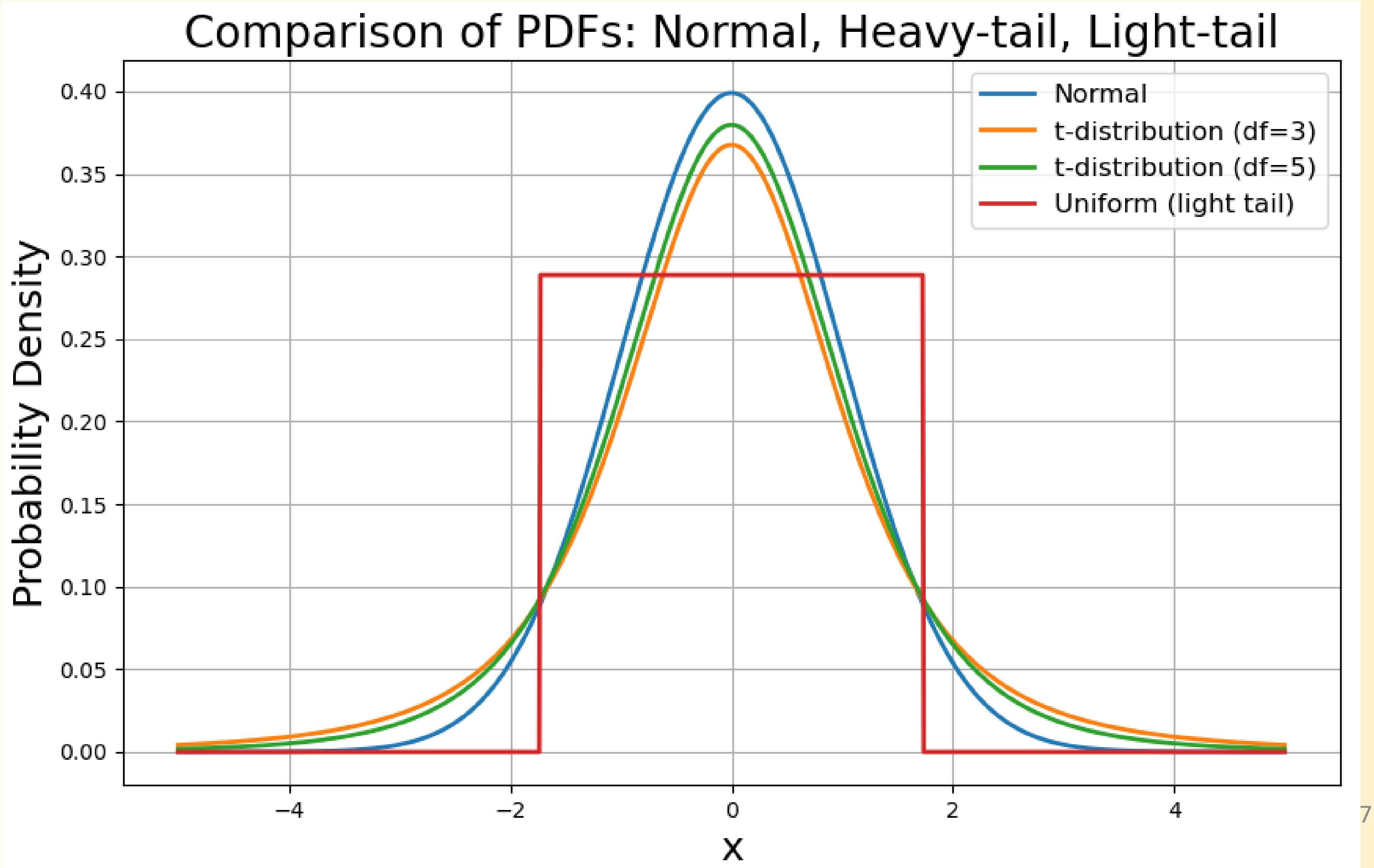
$$\beta = P(485 \le \bar{x} \le 515 \text{ when } \mu = 520)$$



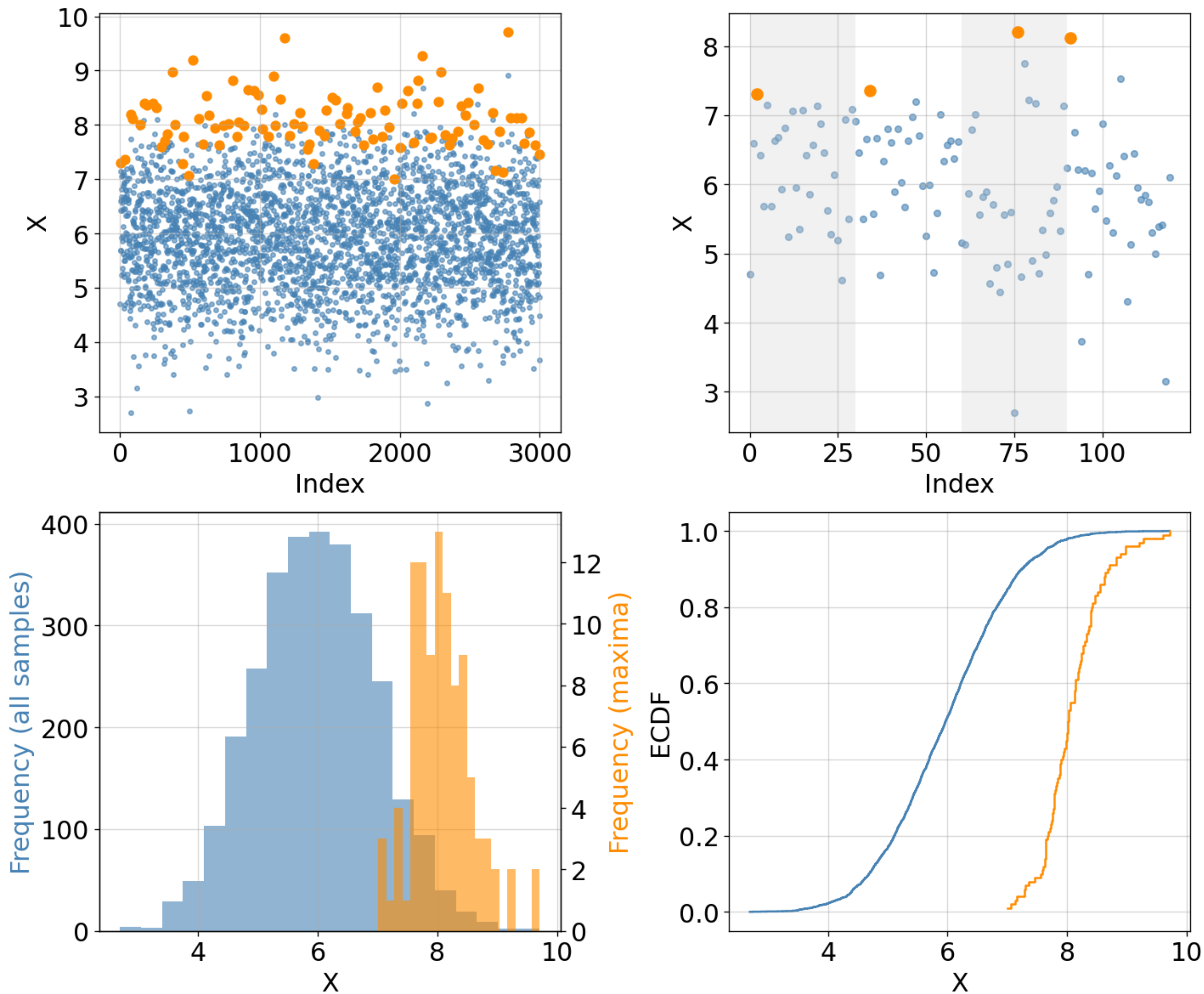
Extreme Value Analysis

Tail of distributions

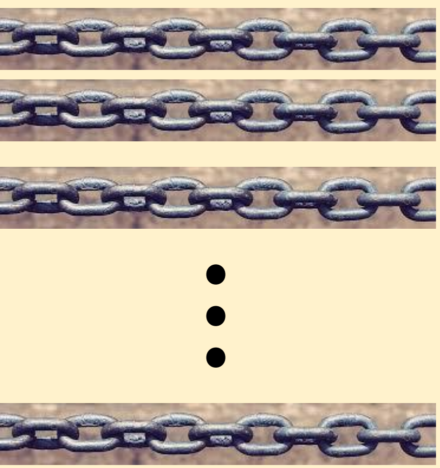
- *Excess Kurtosis:* $E[(X-\mu)^4]/\sigma^4 - 3$

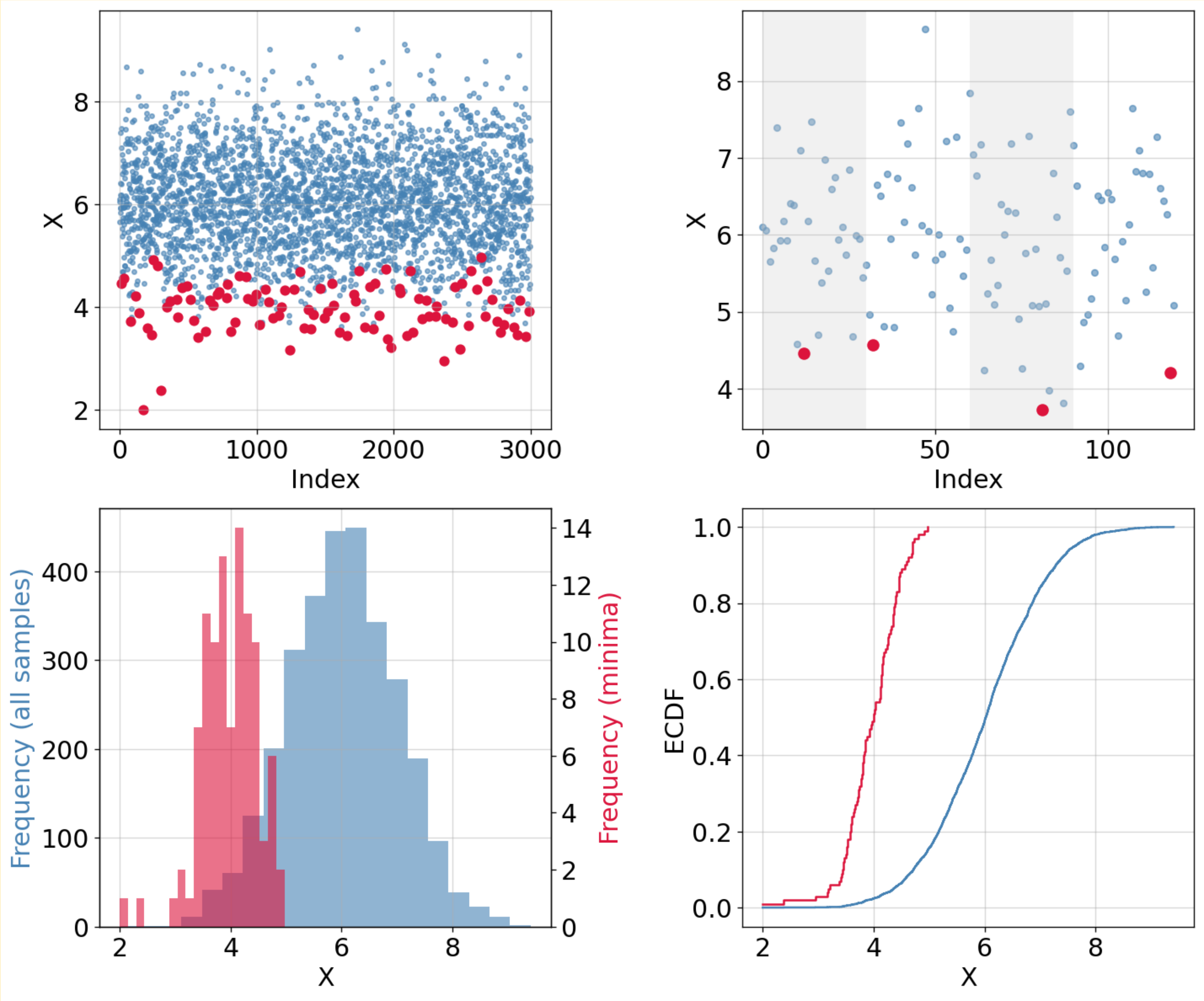


Extreme value analysis

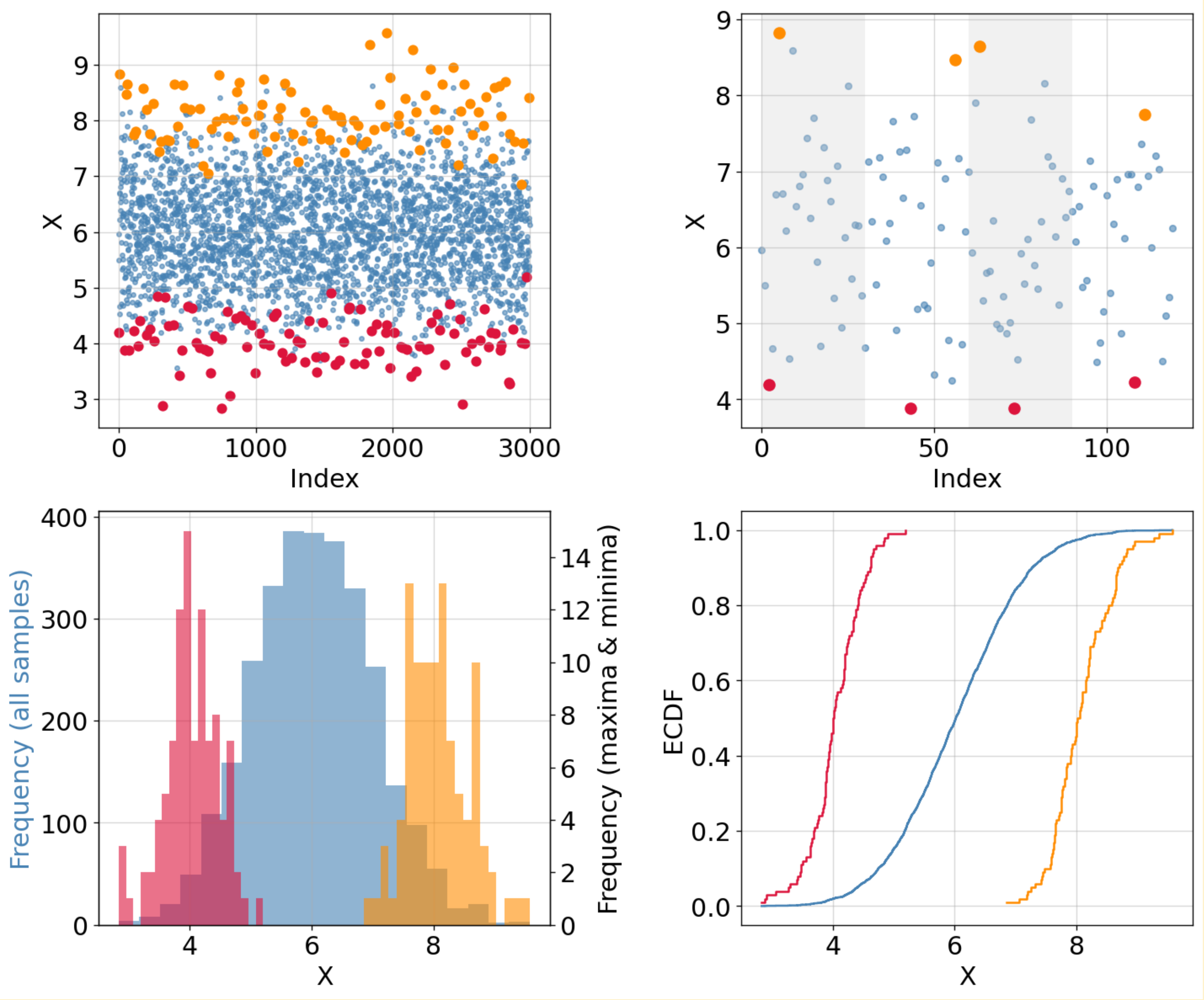


Extreme value analysis



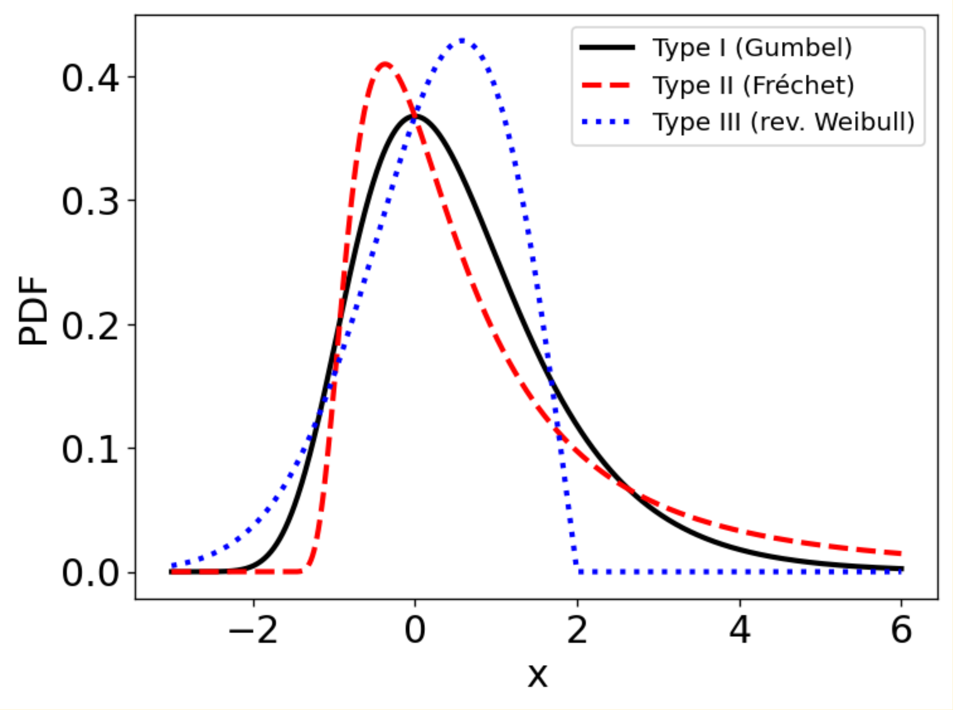


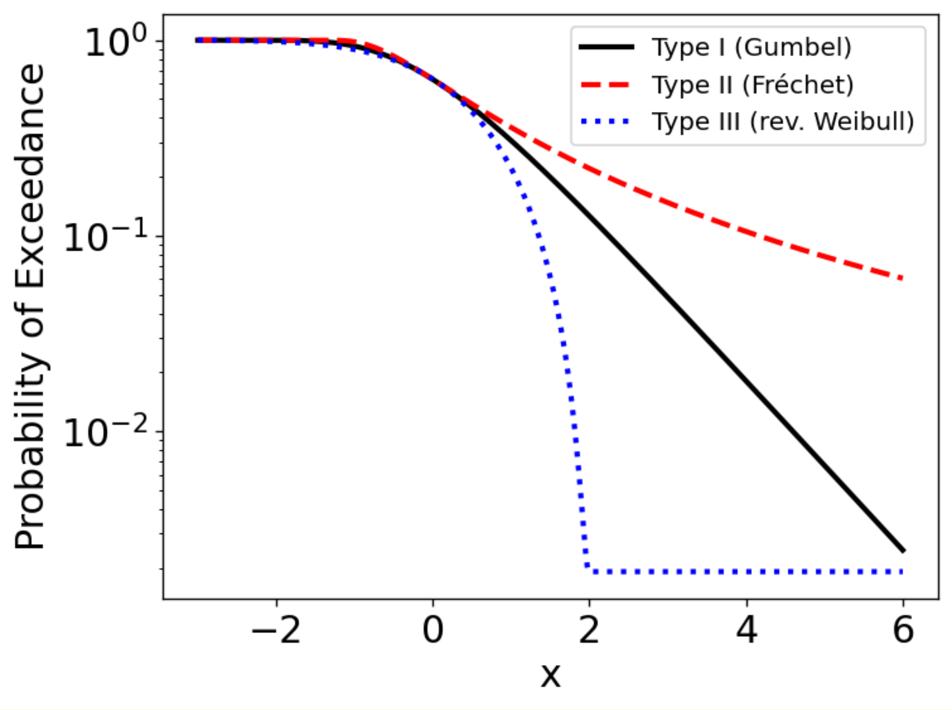
Extreme value analysis



Selection of GEV distribution type

Tail type	Extreme value type	Parent distribution
Medium-/baseline tailed	Gumbel	Normal
Heavy-/fat-tailed	Fréchet	t-distribution
Light-/thin-tailed	Reversed Weibull	Uniform, beta





Extreme value analysis

- Magnitude of extremes

  Generalized extreme value (GEV) distributions

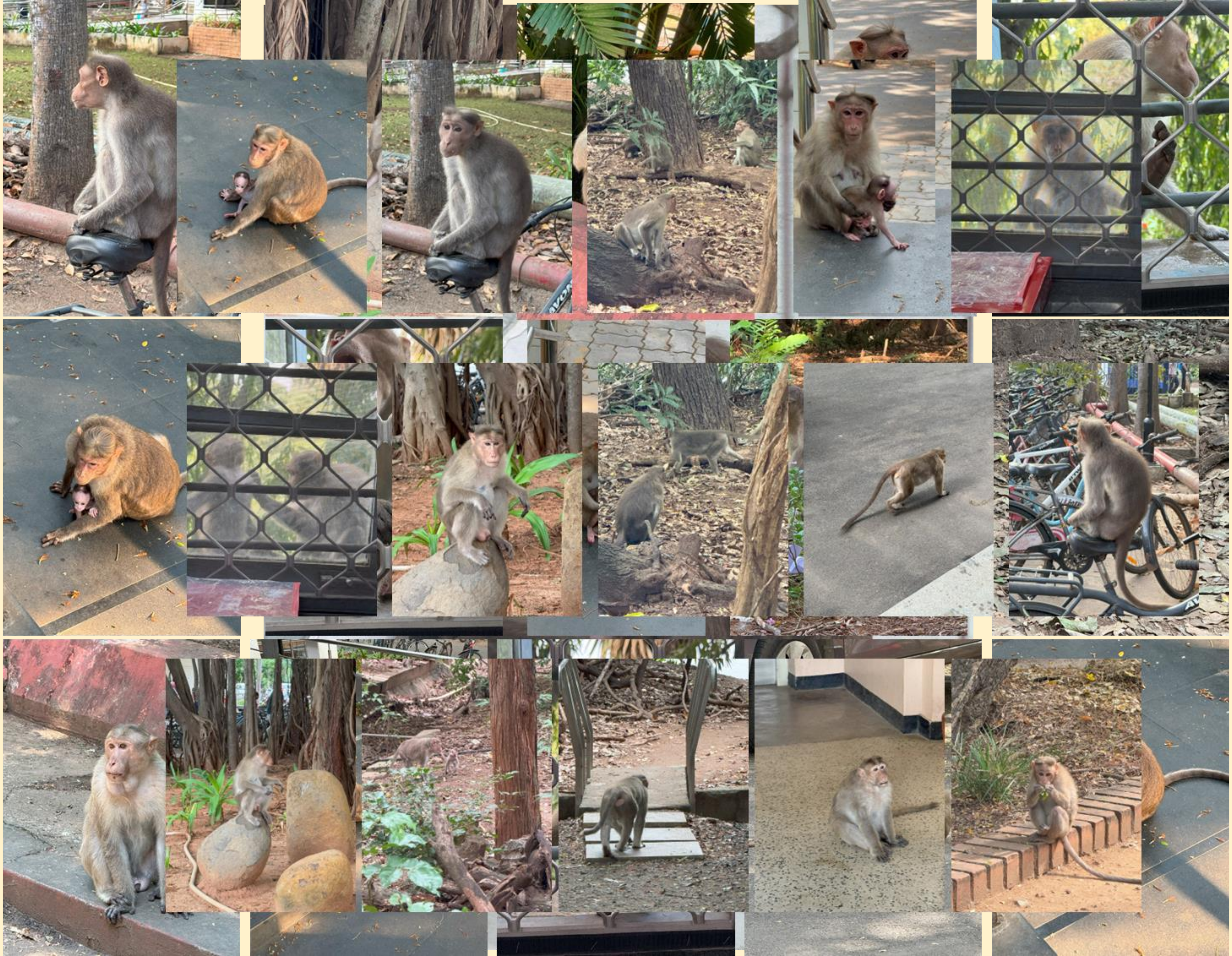
  - Smallest/largest values?
  - Look out for the tails? Thin/thick? Bounded?
  - Estimate excess Kurtosis
  - Pick one of the GEV models

- Frequency of extremes

  Poisson process

  - Count of extremes → Estimate event rate (λ)
  - Independent or clustered?
    - Proceed with Poisson models
    - Decluster using peak-over-threshold/block maxima

Risk & Reliability



Risk from monkey attack on campus

- The risk from the monkey hazard involves:

(or lack thereof)

- Person's familiarity with the hazard (immeasurable; Not useful)
- Severity of attack: How bad was the attack? (intensity measure)

charged scratched bitten mauled

Let's call it "**Monkey Attack Scale, MAS**".

- Rate of monkey attack: How often? (hazard rate)
- Damage to person: Time in the hospital? (damage measure)

None Hours days week-or-more

none minor major severe

- Consequence of damage: $, downtime? (consequence function)

Probabilistic seismic risk assessment

- Three components

Given: location & building design ($\mathcal{L}$ and $\mathcal{D}$)

- Seismic hazard, $\mathcal{H}_{sa} \coloneqq k_0 s_a^{-k}$
  - Occurrence rate, $g(S_a|\mathcal{L})$ Depends on location/seismicity
- Building's fragility, drift demand, $\mathcal{F}_{col,\mathcal{D}}(s_a) \coloneqq \Pr(col|s_a, \mathcal{D})$
  - Probability density, $p(DM|S_a, \mathcal{L}, \mathcal{D})$ Depends on building (type, material, age)
- Consequence of damage, e.g., downtime, repair cost ratio:

  $\mathcal{C}_{DT}(dm) \coloneqq \Pr(DT = dt|dm)$ or $\mathcal{C}_{RCR}(dm) \coloneqq \Pr(RCR = rcr|dm)$
  - Probability density, $p(DV|DM, \mathcal{L}, \mathcal{D})$

$$g[DV|\mathcal{L}, \mathcal{D}] = \iint p(DV|DM, \mathcal{L}, \mathcal{D})\, p(DM|IM, \mathcal{L}, \mathcal{D})\, g(IM|\mathcal{L})\mathrm{d}IM\ \mathrm{d}DM$$

$$g[DV|\mathcal{L}, \mathcal{D}] = \iiint p(DV|DM, \mathcal{L}, \mathcal{D}) p(DM|EDP, \mathcal{L}, \mathcal{D}) p(EDP|IM, \mathcal{L}, \mathcal{D}) g(IM|\mathcal{L})\mathrm{d}IM\ \mathrm{d}DM$$

Questions, comments, or concerns?