International Association of Oil & Gas Producers

virtual engagement session for OSRC 2021
fixed offshore structures

pre-meeting video 1
basic probability & statistics

pre-meeting background on…

- basic probability and statistics
- structural probability and statistics
- metocean probability and statistics

content – notation & terminology

- notation – typically compressed
- discreet variable v continuous variable
- probability mass function v probability density function
- random variable v deterministic variable
- parameters v variables
- independently and identically distributed (iid)
- addition rule, multiplication rule & chain rule
- parent v extreme distributions
- marginal probability, conditional probability, Bayes' rule & Bayesian inference

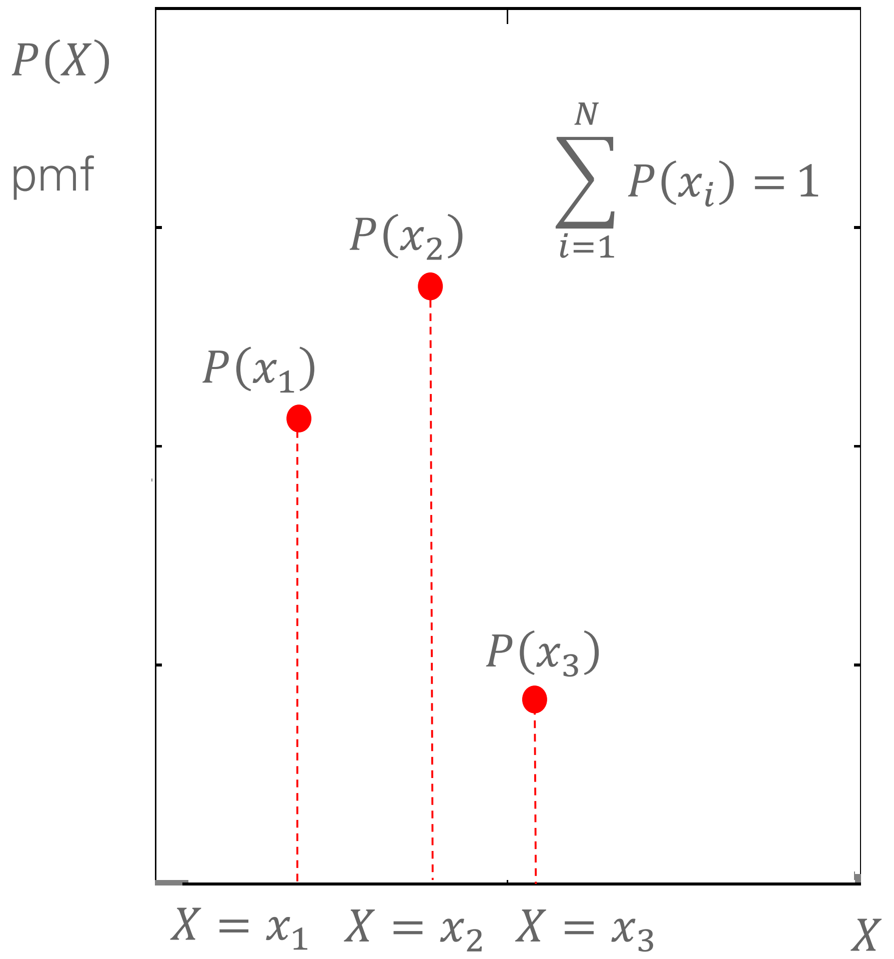
probability – definition

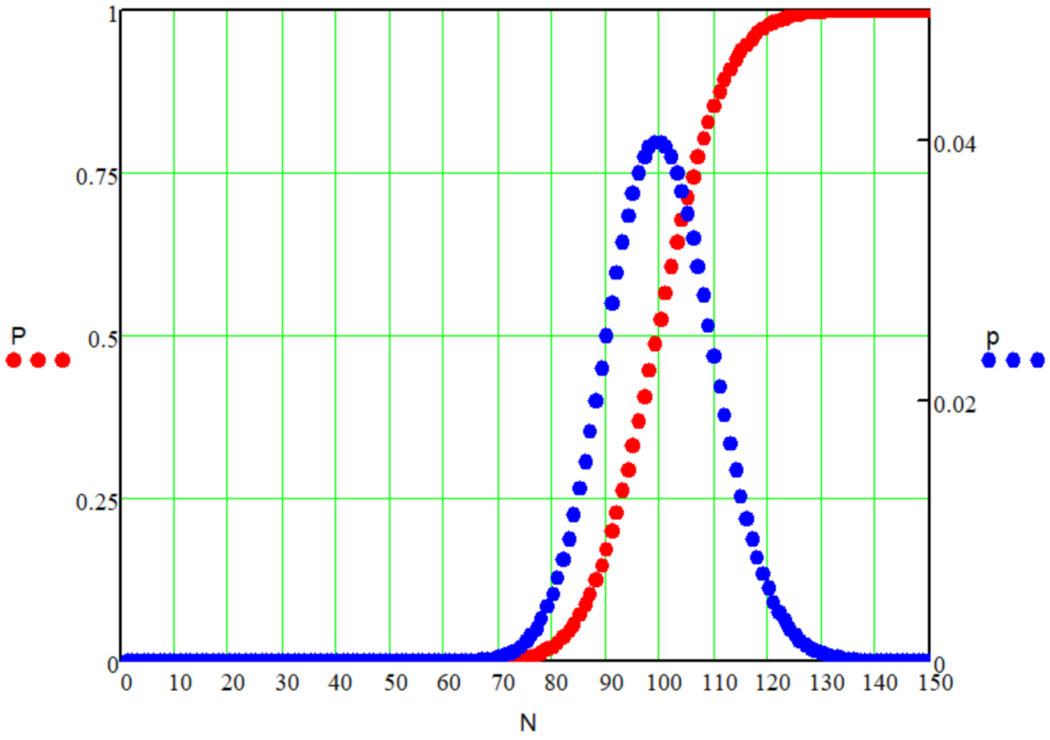
probabilities, P ,are numerical quantities…

- defined on a set of "outcomes" ....eg $P(C \leq 20m),\ P(20 < C \leq 22m),\ P(C > 22m)$
- non-negative
- additive over mutually exclusive outcomes ....eg $P(C \leq 20m) + P(20 < C \leq 22m)$
- sum to 1 over all possible mutually exclusive outcomes

probability of an event A $= P(A) = \dfrac{\text{number of ways event A can occur}}{\text{total number of possible outcomes}}$

probability mass function (pmf) & probability





X– the variable

x_i – a specific value of the variable

probability density function (PDF) – in 1 dimension

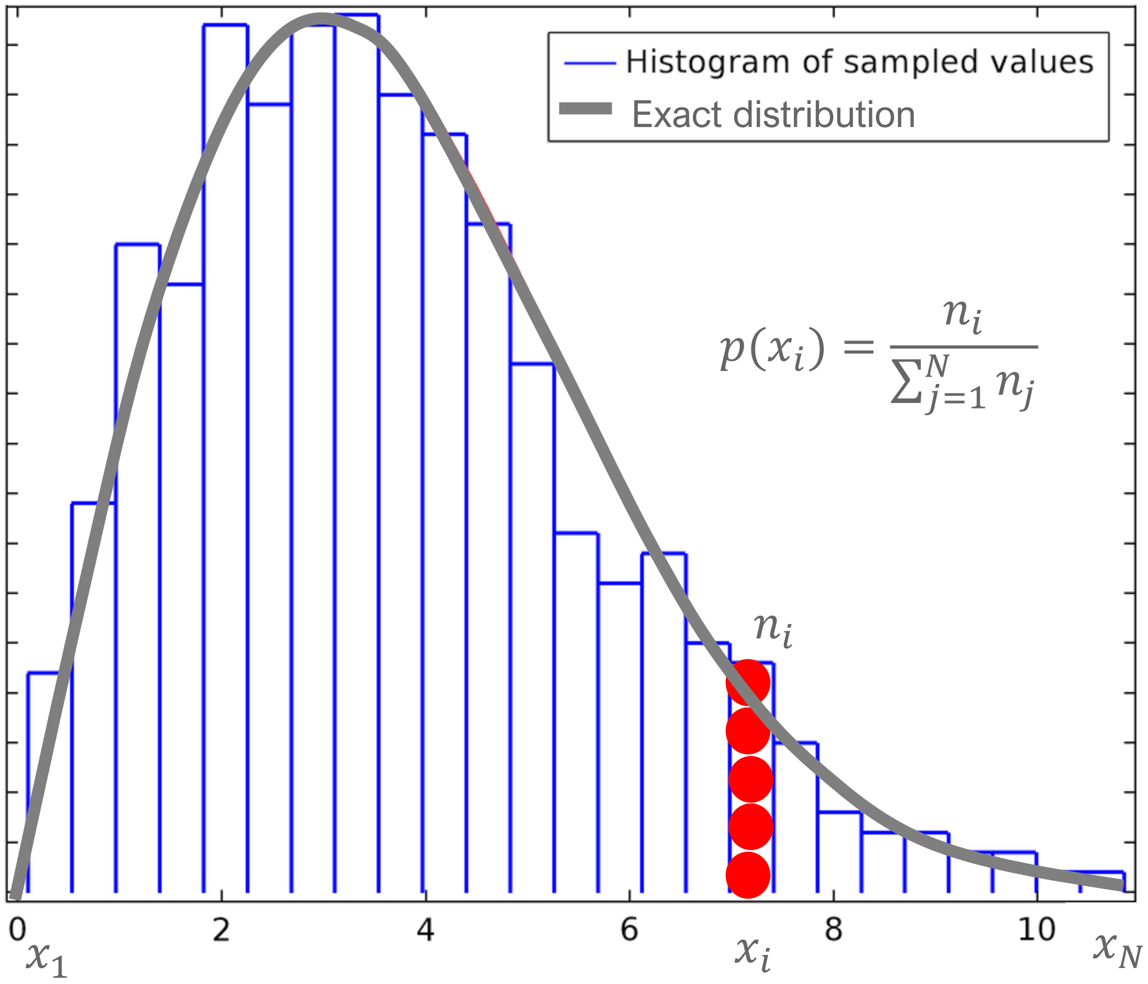
X– the variable

x – a specific value of the variable

Francis Galton's 1894 "bean machine" for simulating normal distributions

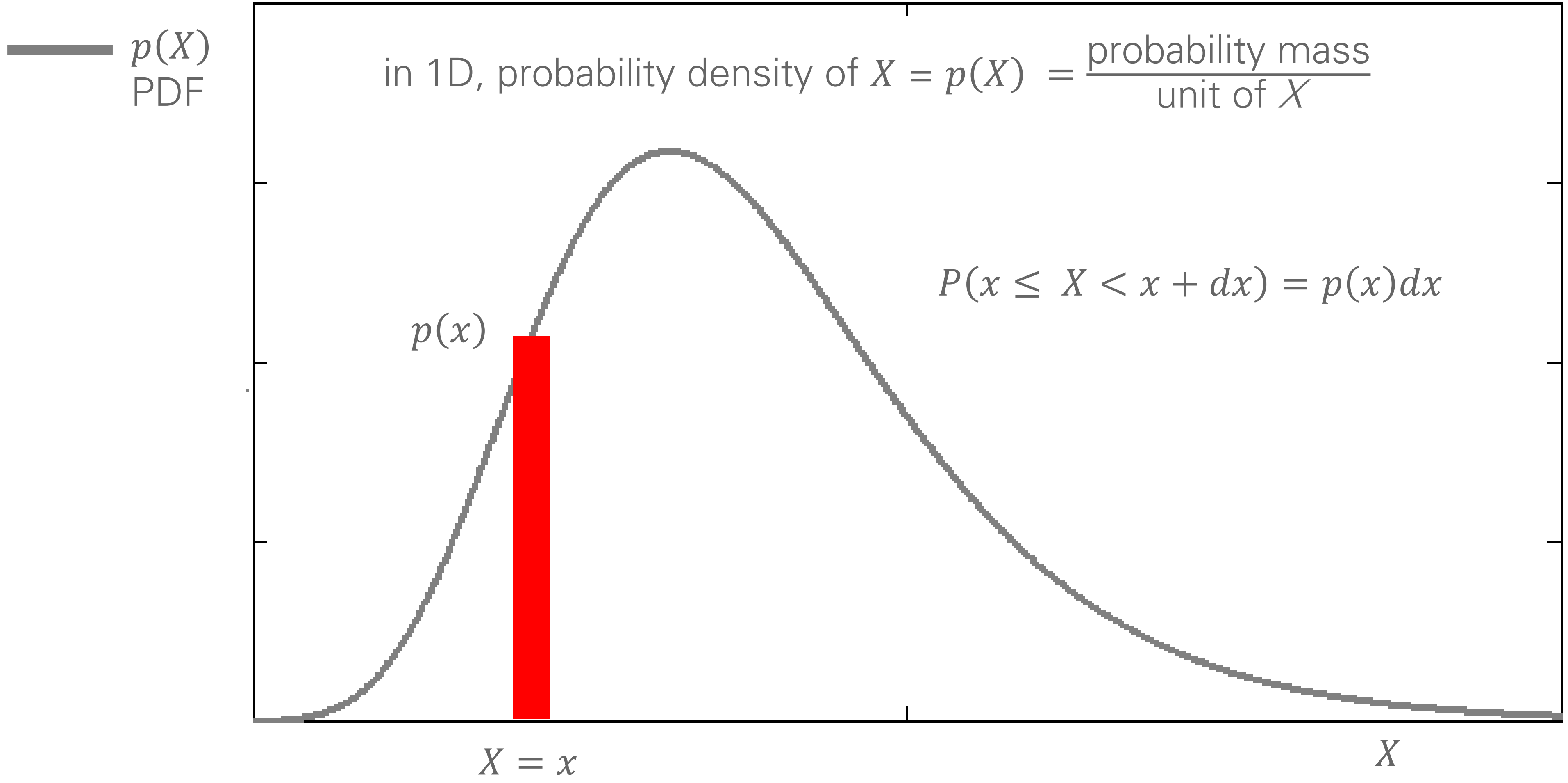
$P(x_i < X < x_{i+1})$

$p(X)$ PDF

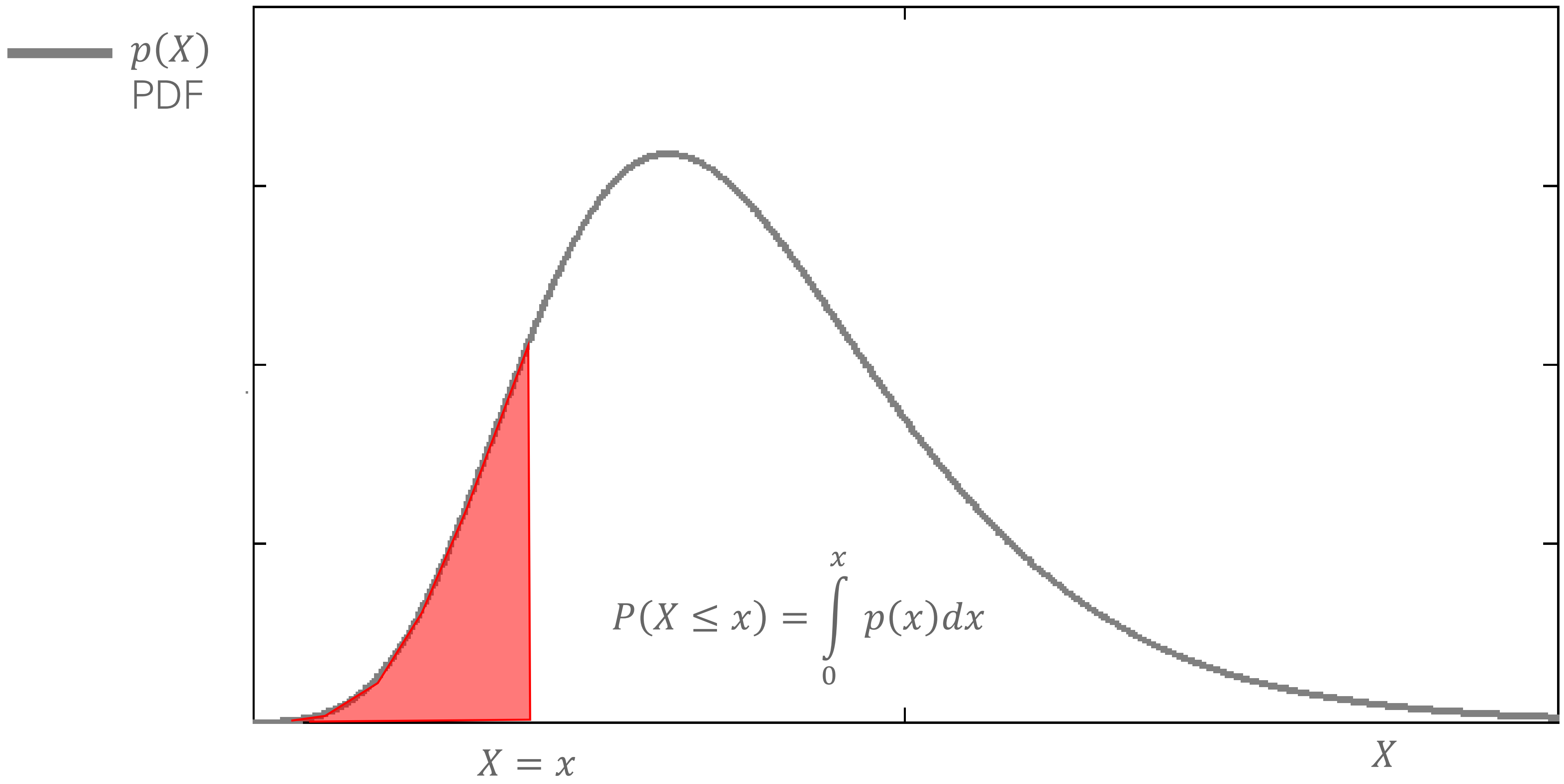


physical quantities that are expected to be the sum of many independent processes often have distributions that are nearly normal (central limit theorem)

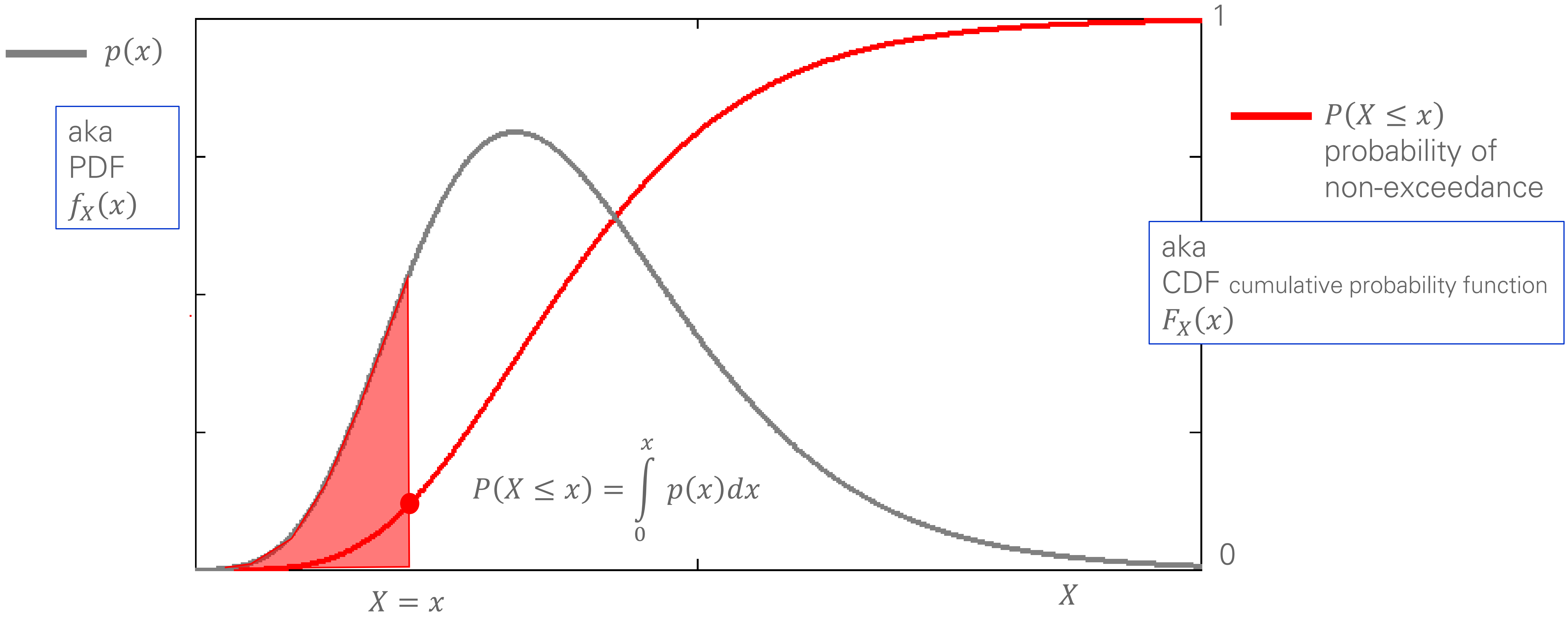
probability density & probability



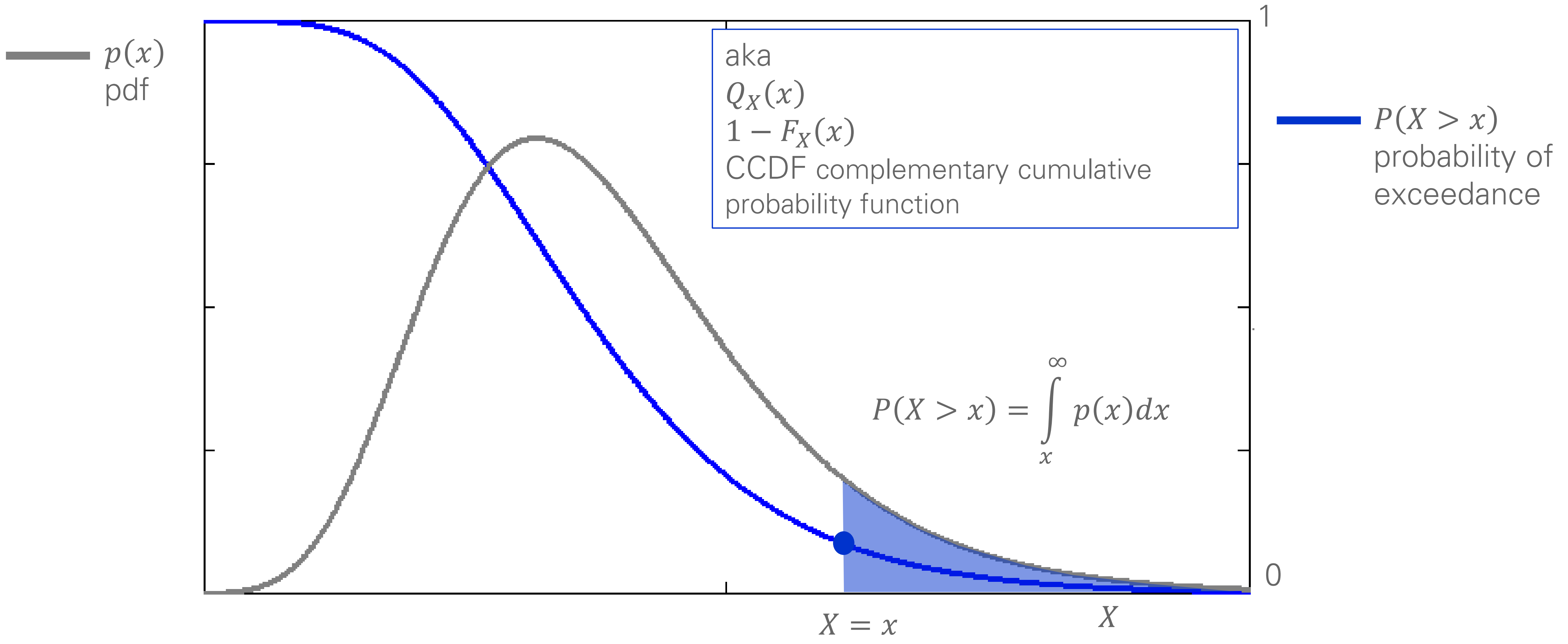
probability density & probability of non-exceedance



probability density & probability of non-exceedance



probability density & probability of exceedance

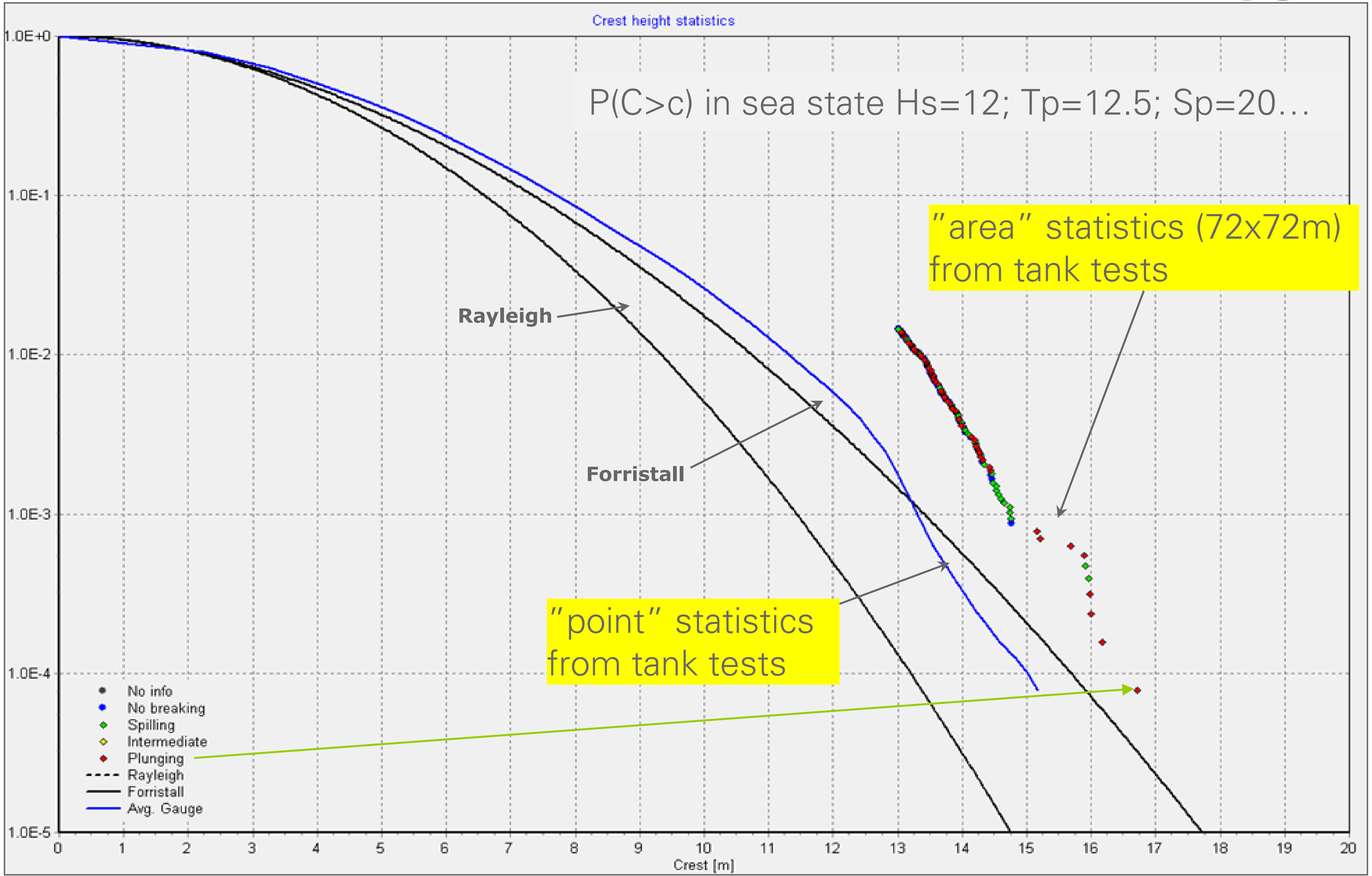


probability of exceedance

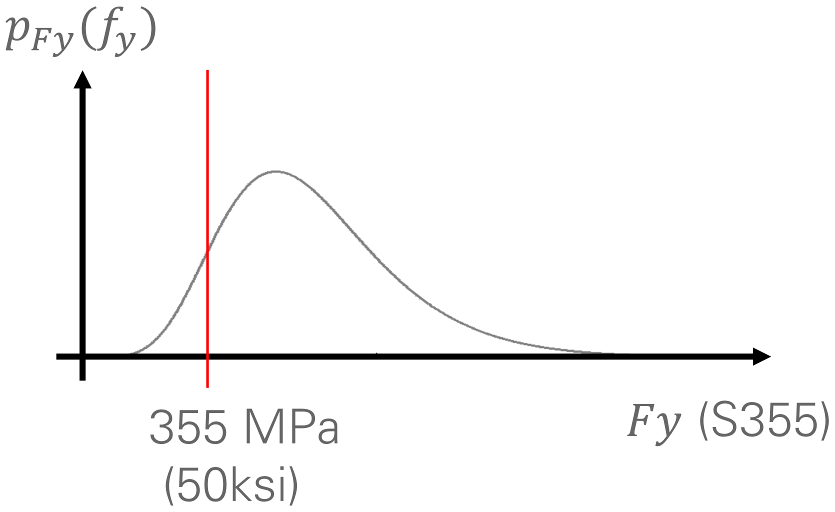
$P(C > c)$

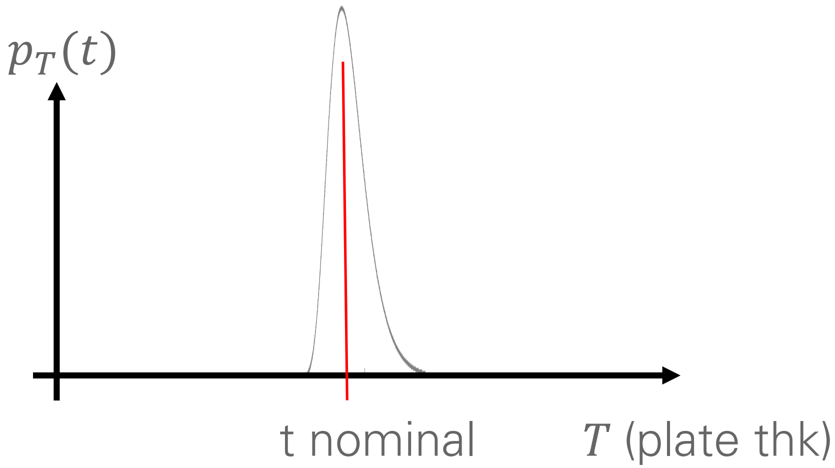
probability of exceedance of individual crest ht (C) in a given sea state

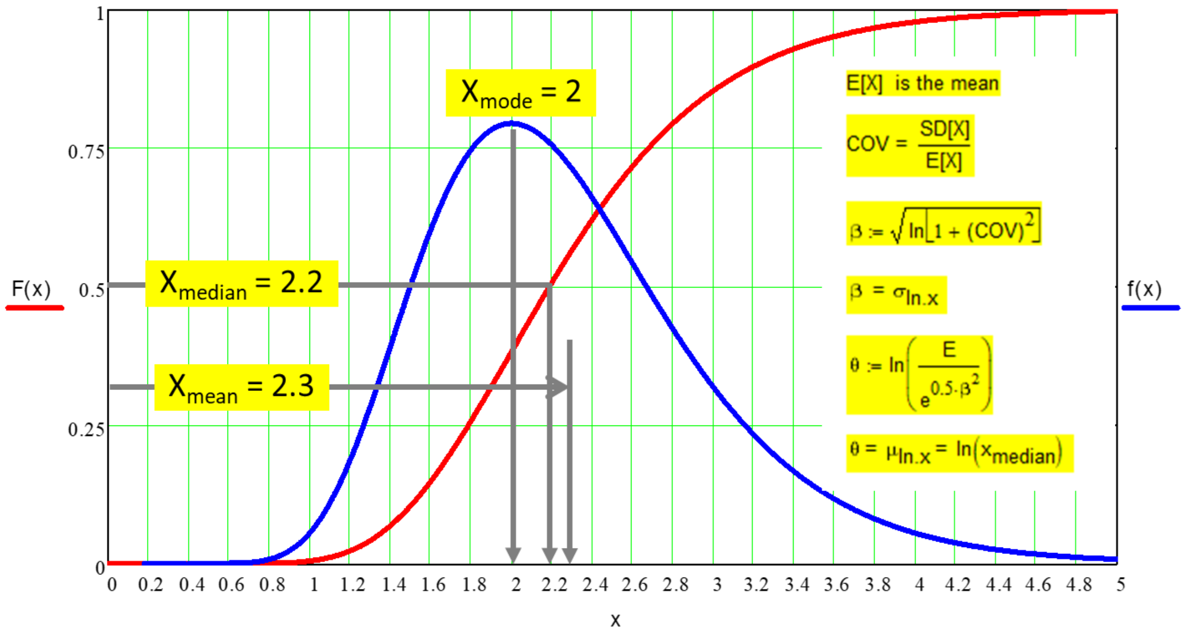
log scale rather than linear (0 to 1) shows the tail in more detail at extreme values



random variables (v deterministic variables)







parameters

location μ
mean (expectation)
mode (mp)
median (P50)

scale
standard deviation σ
variance σ^2
COV σ/μ
dispersion $\beta = \sigma_{logx}$

shape ξ
tail properties

standard probability density functions

Continuous distributions

Uniform
Normal
Lognormal
Gamma
Inverse-gamma
Chi-square
Inverse-chi-square
Scaled inverse-chi-square
Exponential
Laplace
Weibull
Wishart
Inverse-Wishart
LKJ correlation
t
Beta
Dirichlet
Logistic
Log-logistic

Discrete distributions

Poisson
Binomial
Negative-binomial
Beta-binomial

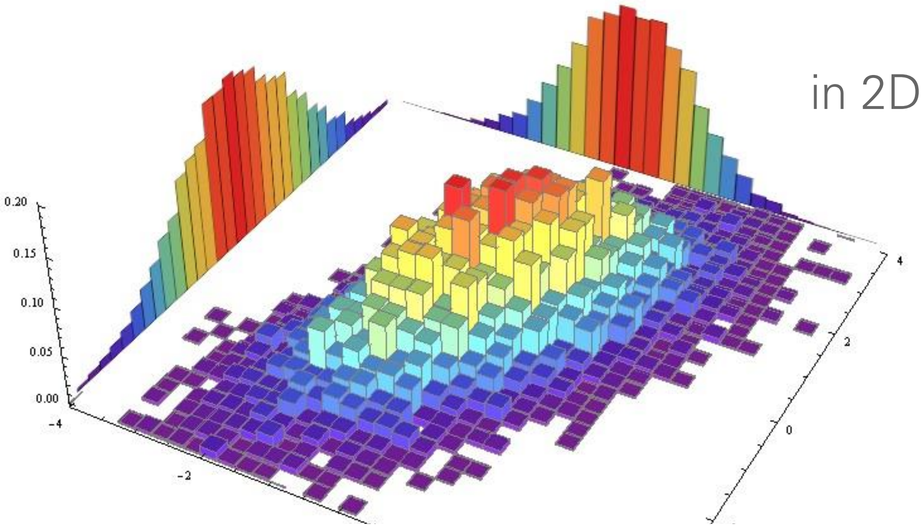
$$p(x) = \frac{1}{b-a}$$

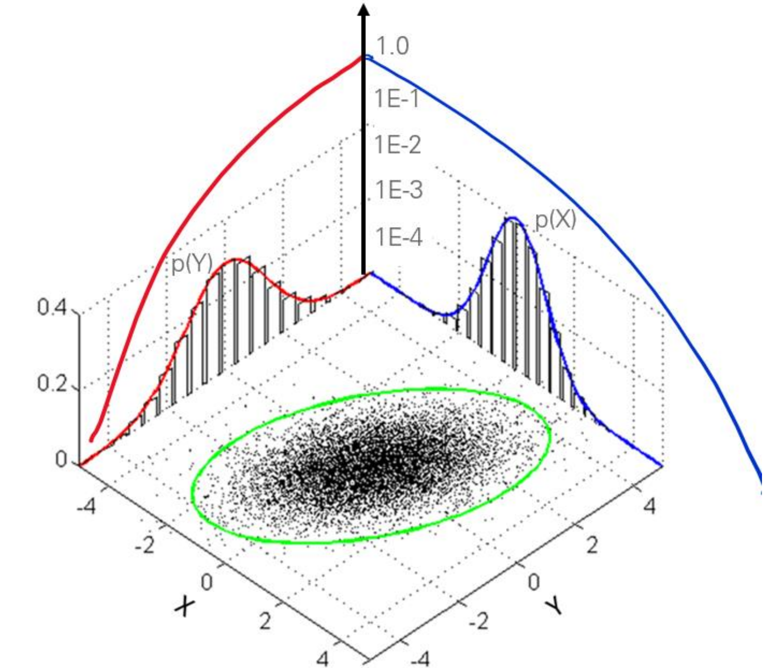
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

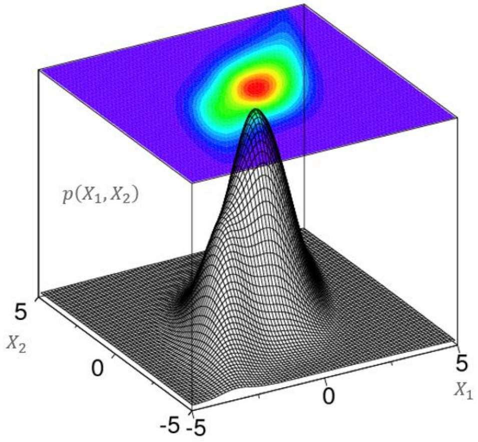
$$p(x) = \frac{1}{\sigma x\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(\log x-\mu)^2\right)$$

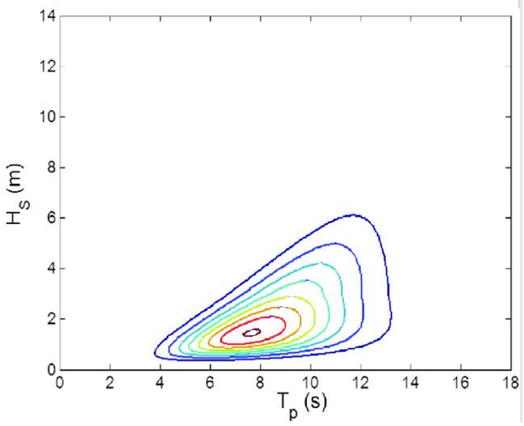
joint probability density function - 2D

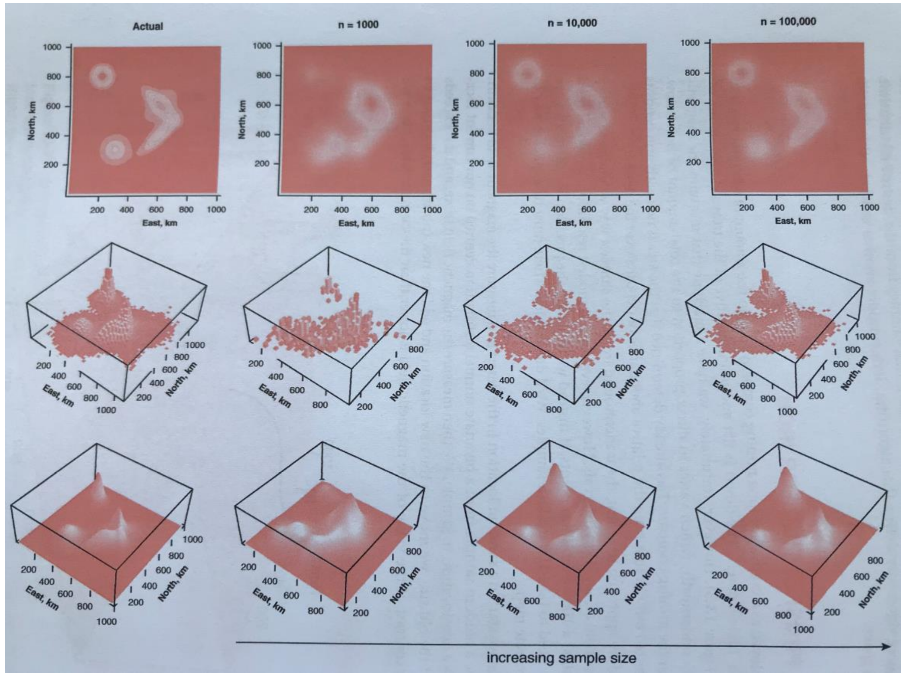
in 2D, probability density $= p(X,Y) = \dfrac{\text{probability mass}}{\text{unit of X} \times \text{unit of } Y}$





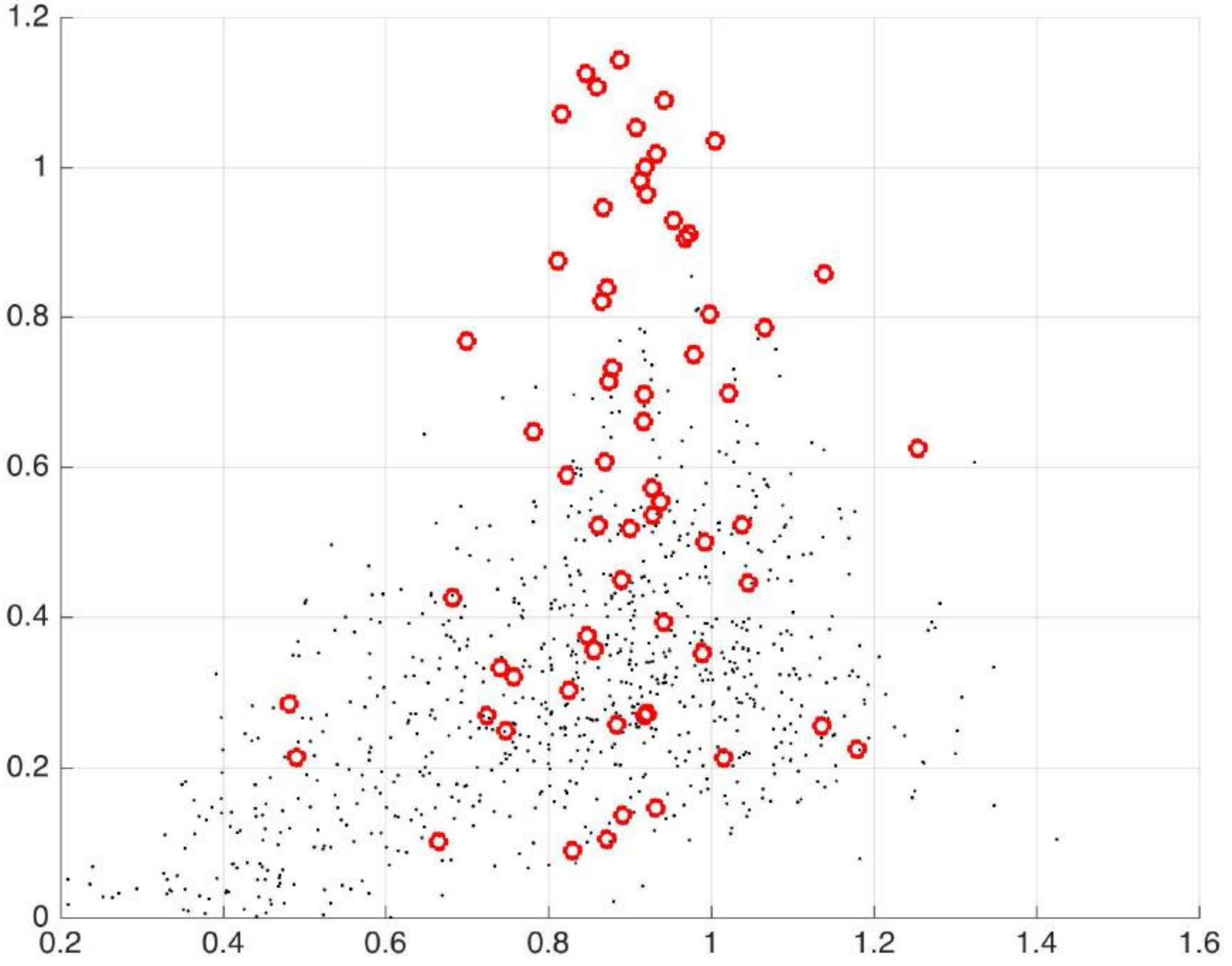






computing the pdf by sampling using MCMC (HMC)

joint probability density function - 2D



$\frac{C}{H_s}$

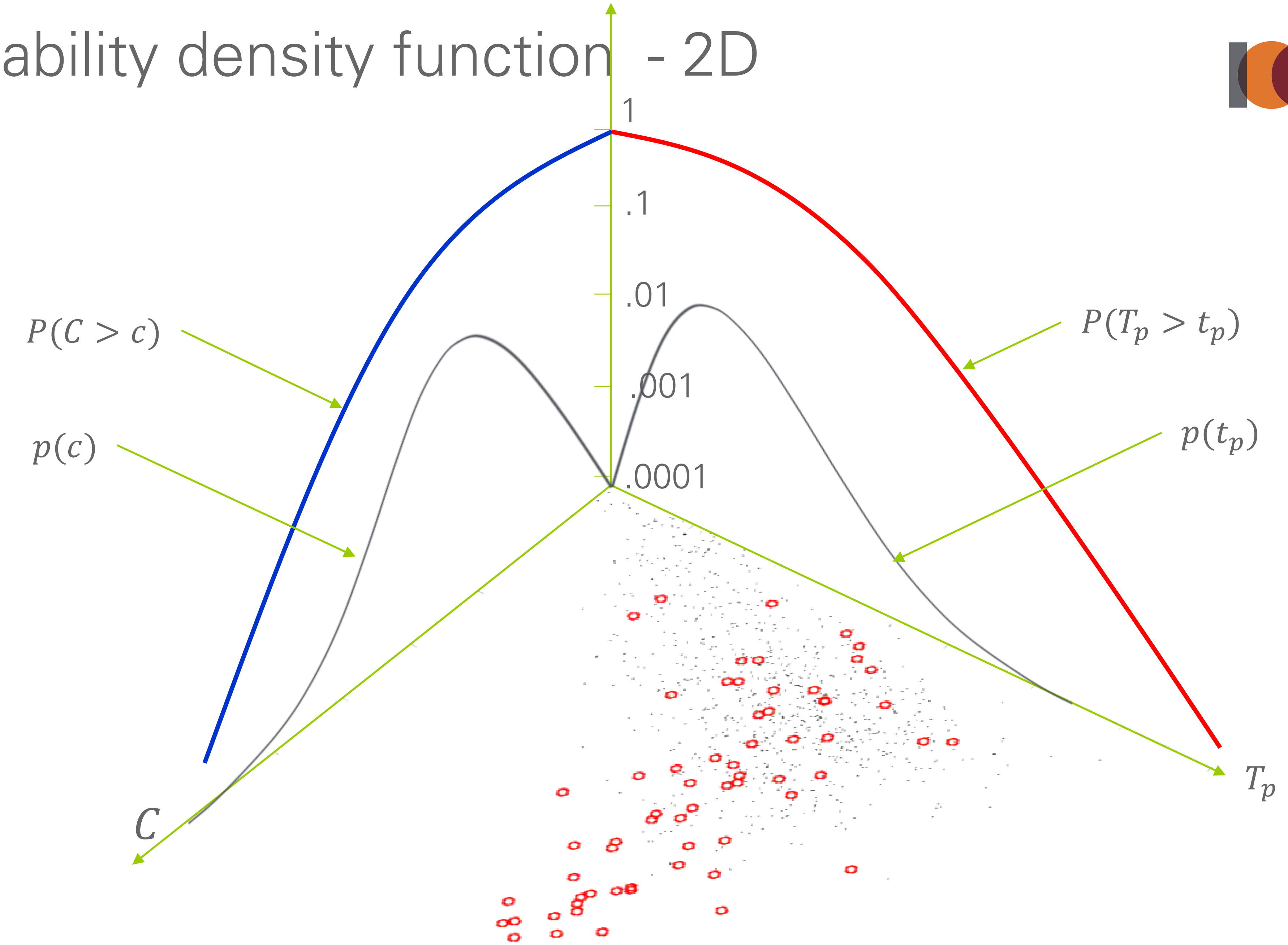
$\frac{T}{T_p}$

scatter plot

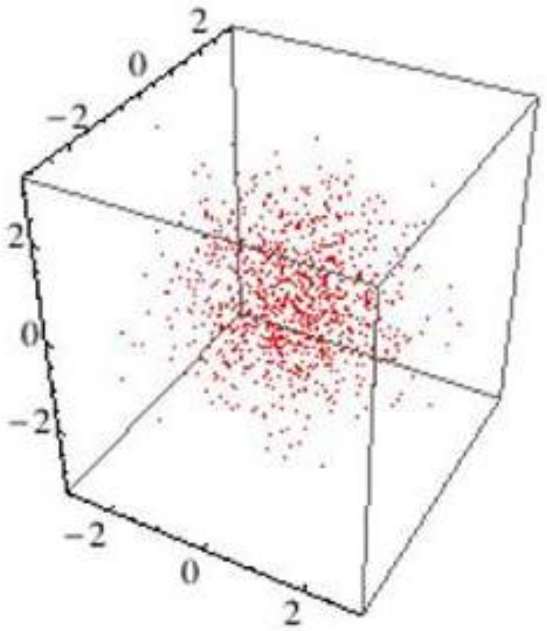
grey dots = Monte Carlo (random) realisations from the joint distribution of (C, T_p)

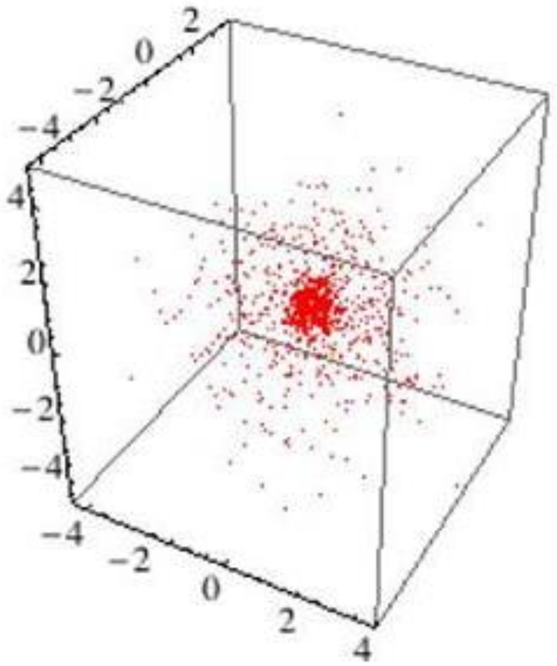
red dots = stratified Monte Carlo realisations from the joint distribution of (C, T_p)

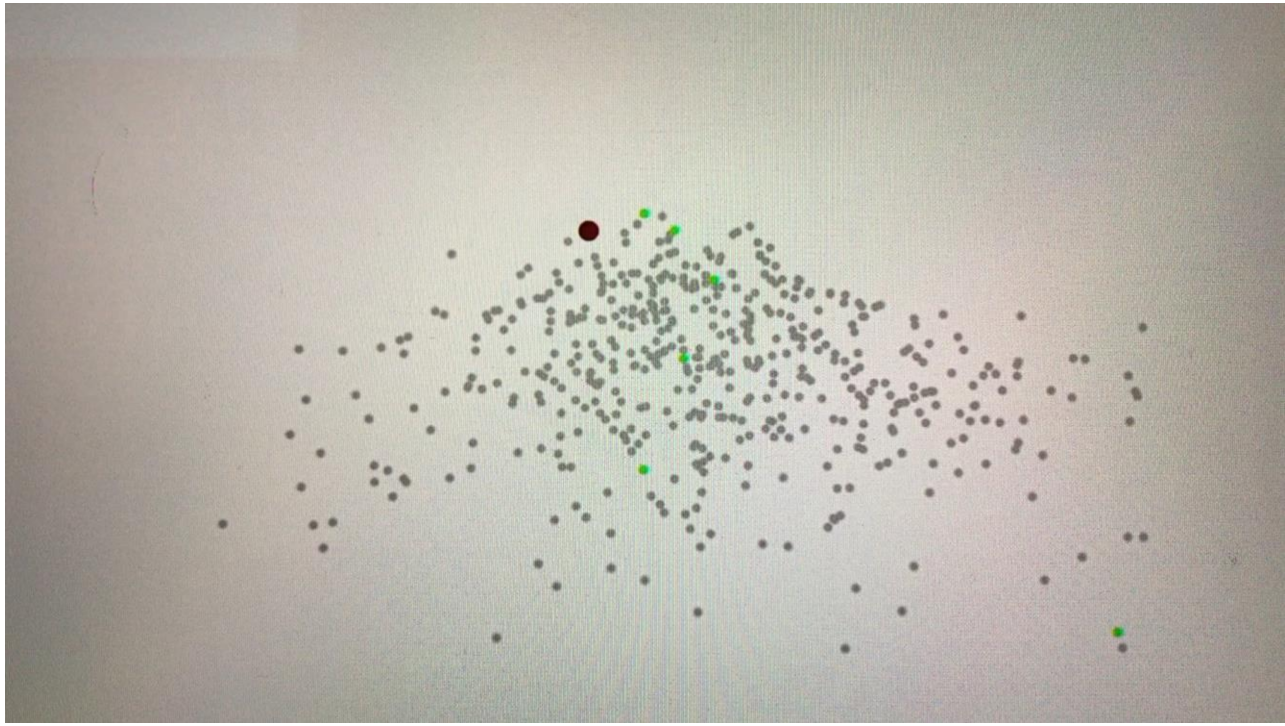
joint probability density function - 2D



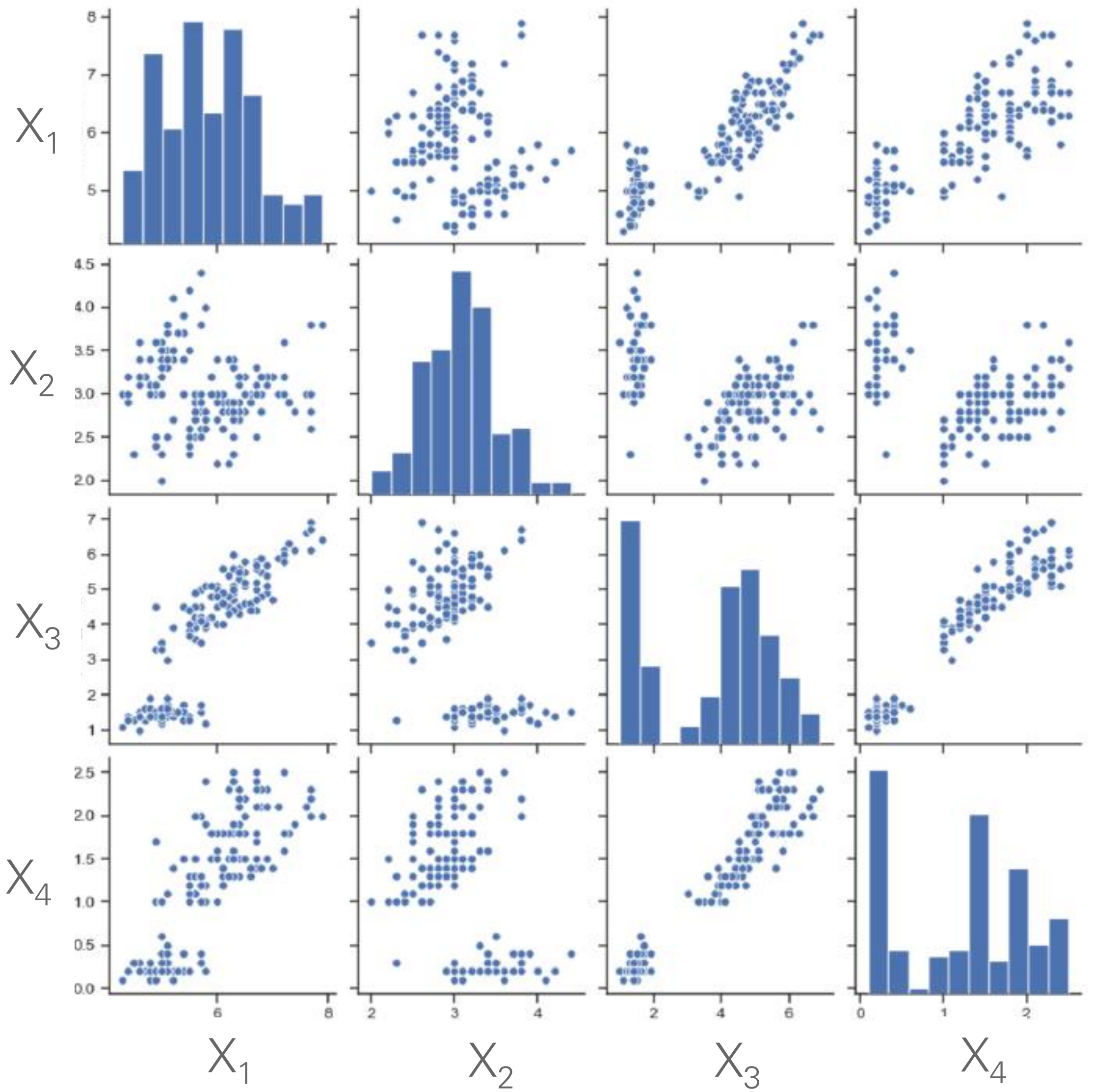
joint probability density function - 3D



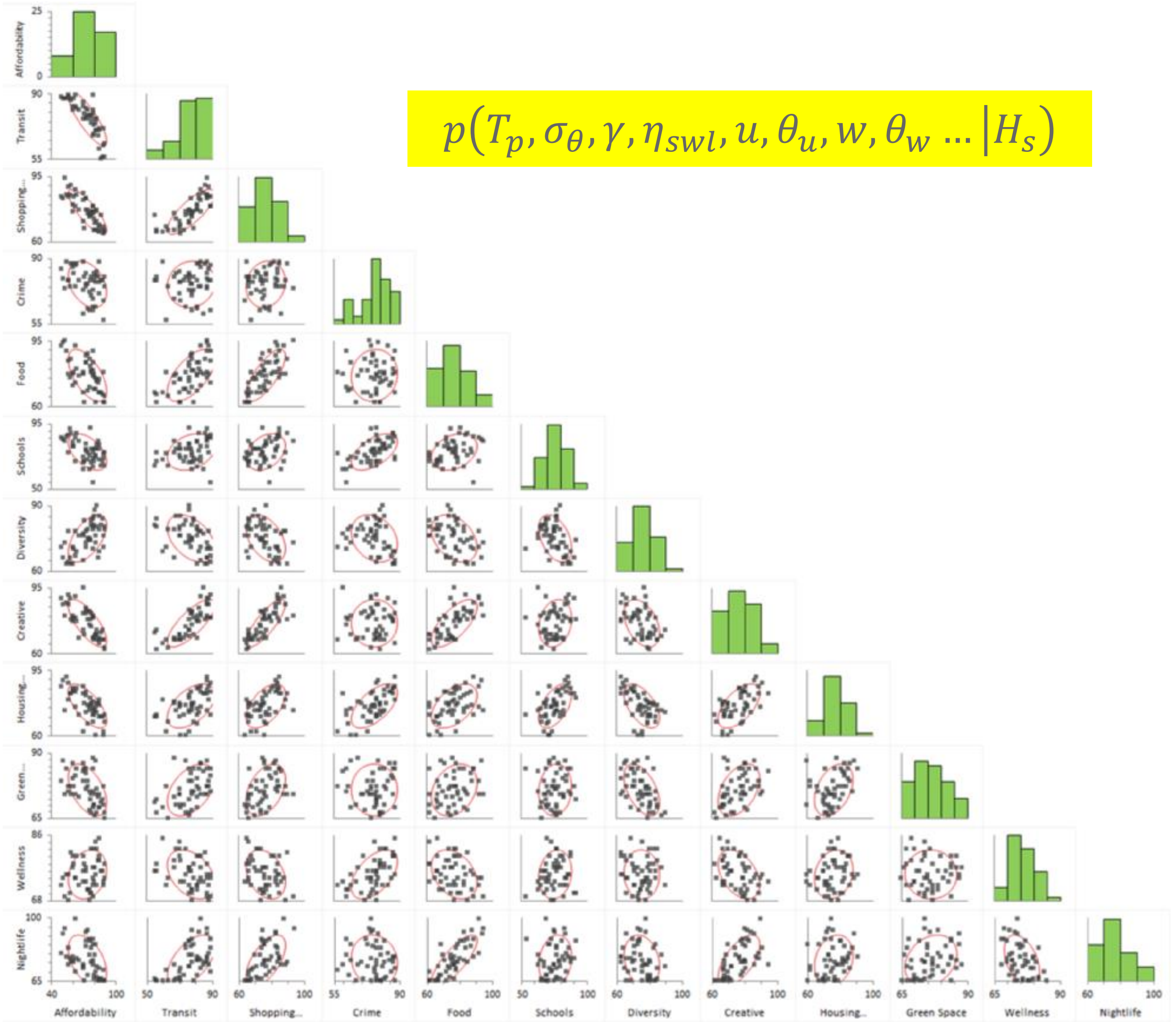




joint probability density function - 4D and 12D



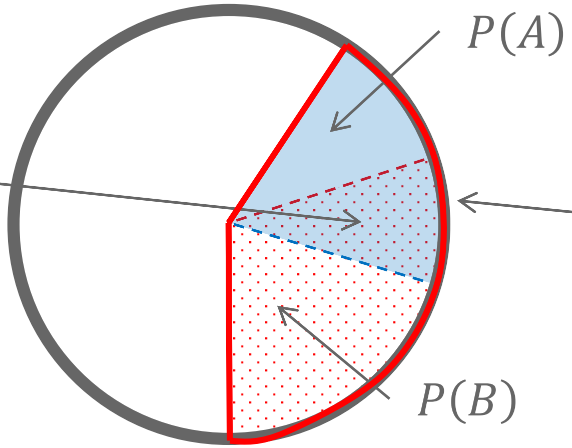
$$p\left(T_p, \sigma_\theta, \gamma, \eta_{swl}, u, \theta_u, w, \theta_w \ldots \middle| H_s\right)$$



addition rule

International Association of Oil & Gas Producers

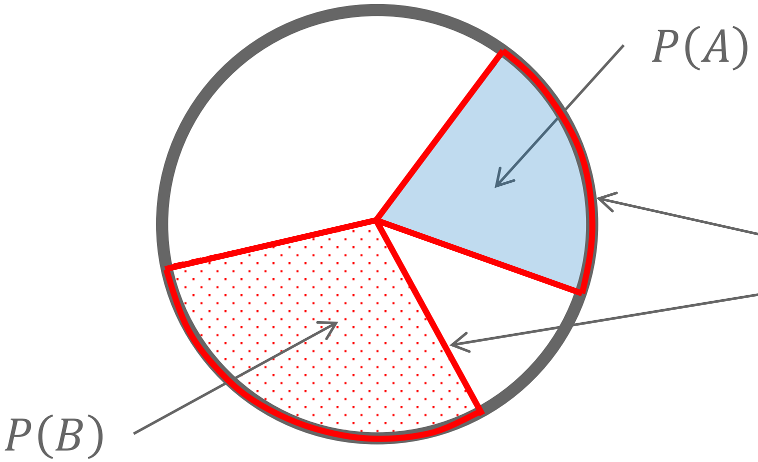
$P(A \text{ AND } B)$
$= P(A \cap B)$
$\neq 0$
= not mutually exclusive



general addition rule

$P(A \text{ OR } B)$
$= P(A \cup B)$
$= P(A) + P(B) - P(A \cap B)$
$= P(A) + P(B) - P(A \text{ AND } B)$

$P(A \text{ AND } B)$
$= P(A \cap B)$
$= 0$
= mutually exclusive



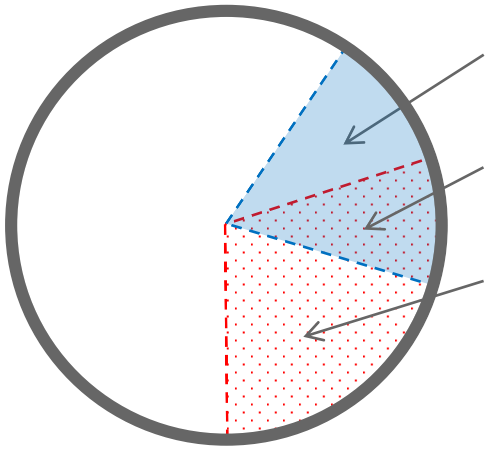
specific addition rule (for mee)

$P(A \text{ OR } B)$
$= P(A \cup B)$
$= P(A) + P(B)$

Venn-pie (pie area=1)

conditional probability

$P(A|B)$ = probability of event A (eg $2 < X < 4$) occurring given event B (eg $0 < X < 3$) has occurred



Venn-pie (pie area=1)

$P\left(E \le \eta \middle| H_{s_i} = h\right)$
probability of wave crest elevation E not exceeding η
given the significant wave height for the i th sea state H_{s_i} equals h

$P(L > l|\alpha, \text{storm})$
probability of jacket base shear load L exceeding l
given a (random) storm from direction α is occurring

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.1}{0.3} = 33\%$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.1}{0.2} = 50\%$$

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A \cap B)}{P(B)} = \frac{\text{intersection}}{\text{normalised}}$$

makes $P(A|B)$ a valid probability

multiplication rule

conditional probability of event A occurring given event B has occurred

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

$$P(A \text{ AND } B) = P(A|B) \times P(B)$$ eqn (1)

$$P(B|A) = \frac{P(B \text{ AND } A)}{P(A)} = \frac{P(B \cap A)}{P(A)}$$

$$P(B \text{ AND } A) = P(B|A) \times P(A)$$ eqn (2)

if event A is independent from event B then

$$P(A|B) = P(A)$$ eqn (3)

$$P(B|A) = P(B)$$ eqn (4)

$$P(A \text{ AND } B) = P(A) \times P(B)$$ from (1) & (3)

$$P(\text{B AND } A) = P(B) \times P(A)$$ from (2) & (4)

Bayes' rule

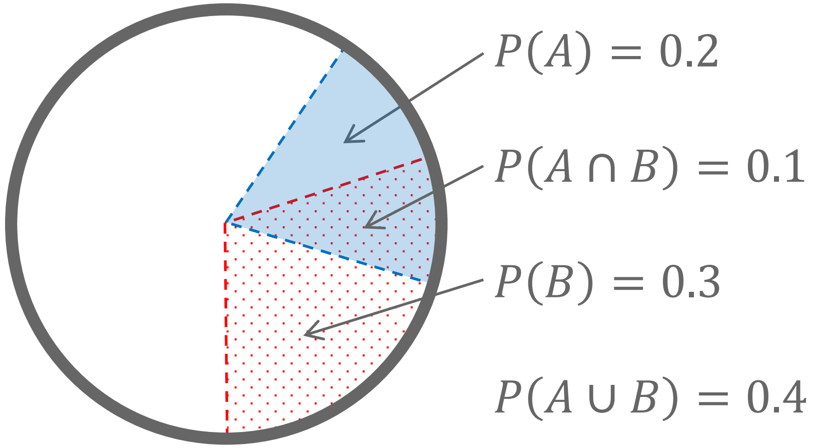
from last slide - as event A and event B occur together then

$$P(A \text{ AND } B) = P(B \text{ AND } A)$$

$$P(A|B) \times P(B) = P(B|A) \times P(A)$$

$$\frac{P(A|B) \times P(B)}{P(B)} = \frac{P(B|A) \times P(A)}{P(B)}$$

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$



Venn-pie (pie area=1)

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.1}{0.3} = 33\%$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.1}{0.2} = 50\%$$

Bayesian inference (probability densities)…

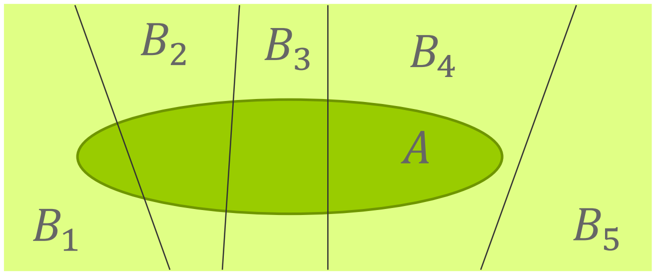
$$p(\boldsymbol{\theta}|y) = \frac{p(y|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})}{p(y)}$$

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} = \frac{0.5 \times 0.2}{0.3} = \frac{0.1}{0.3} = 33\%$$

chain rule

from law of total probability

$$P(A) = \sum_{i=1}^{N} P(A \text{ AND } B) = \sum_{i=1}^{N} P(A \cap B_i)$$



from conditional probability

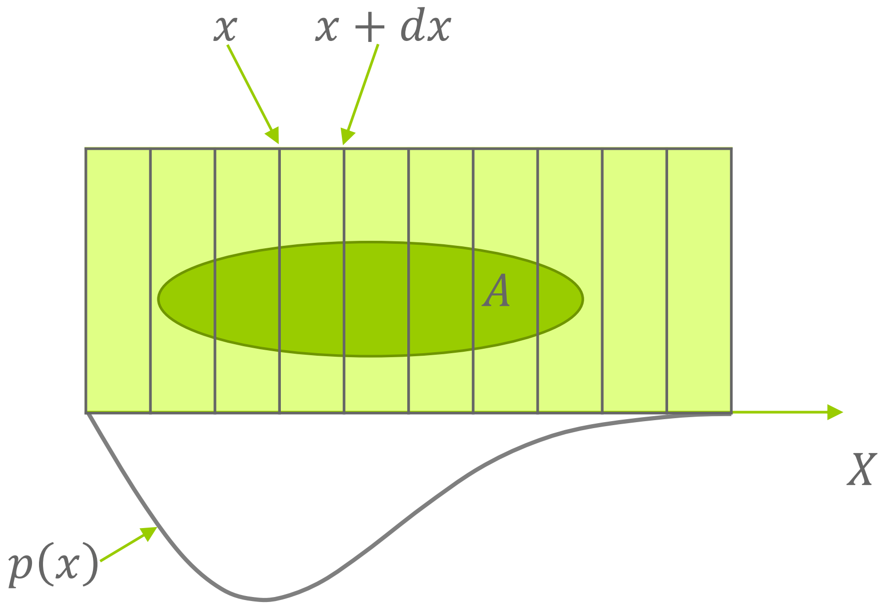
$$P(A \cap B_i) = P(A|B_i) \times P(B_i)$$

substitute 2nd in 1st

$$P(A) = \sum_{i=1}^{N} P(A|B_i) \times P(B_i)$$

extending from discreet events to continuous random variables

$$P(A) = \int_{B} P(A|X = x) \times p(x)\, dx$$



marginal probability

aka marginalising or integrating out

marginal probability density of y

$$p(y) = \int_0^\infty p(y,\theta)d\theta$$

(area =1)

joint probability density of y, θ

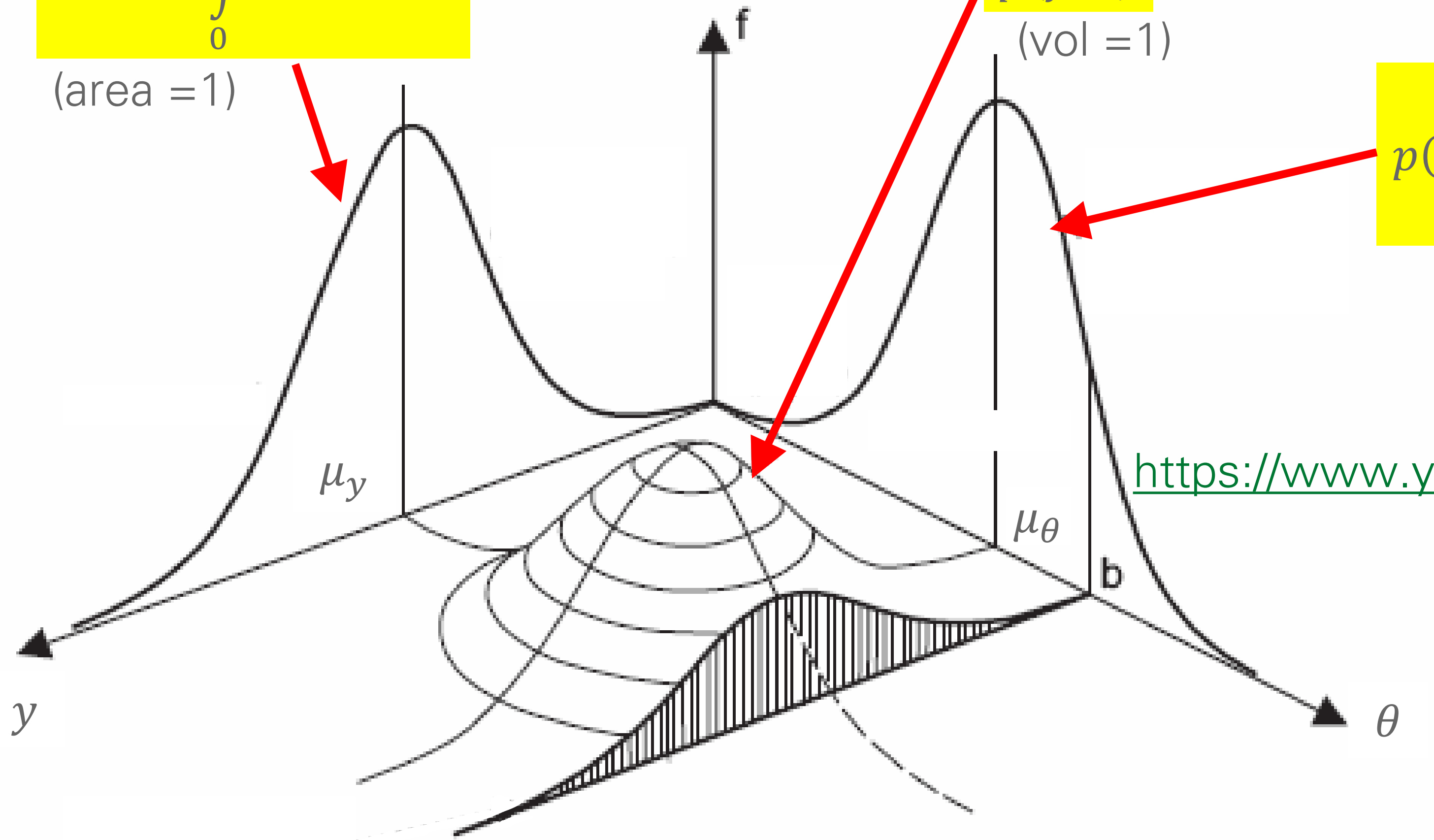
$$p(y,\theta)$$

(vol =1)

$$p(\theta) = \int_0^\infty p(y,\theta)dy$$

(area =1)

marginal probability density of θ



https://www.youtube.com/watch?v=F97Qf5FGt5M

conditional probability

marginal probability density of y

$$p(y) = \int_0^\infty p(y,\theta)d\theta$$

(area =1)

joint probability density of y, θ

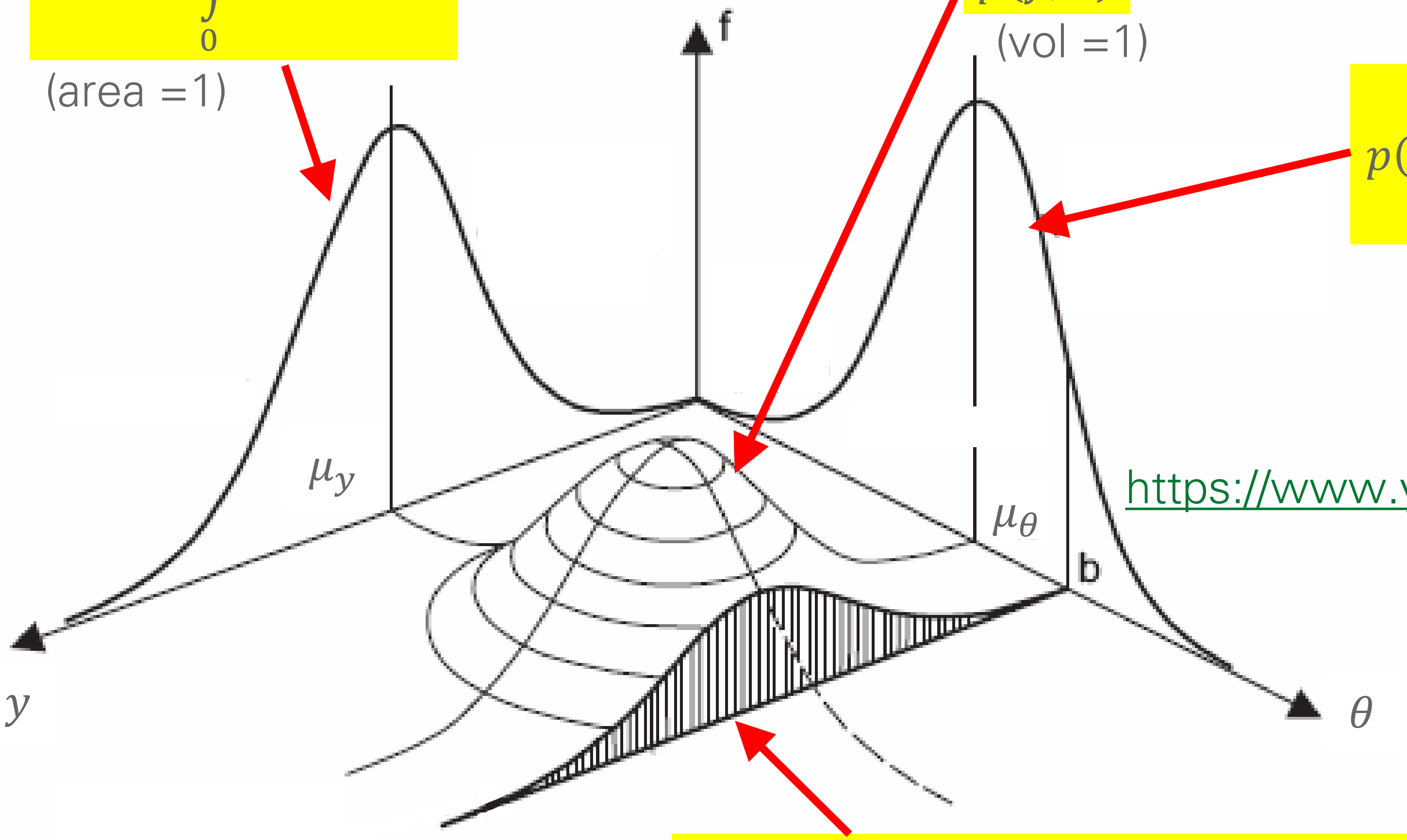
$p(y,\theta)$

(vol =1)

$$p(\theta) = \int_0^\infty p(y,\theta)dy$$

(area =1)

marginal probability density of θ



https://www.youtube.com/watch?v=U7t8lG4E634

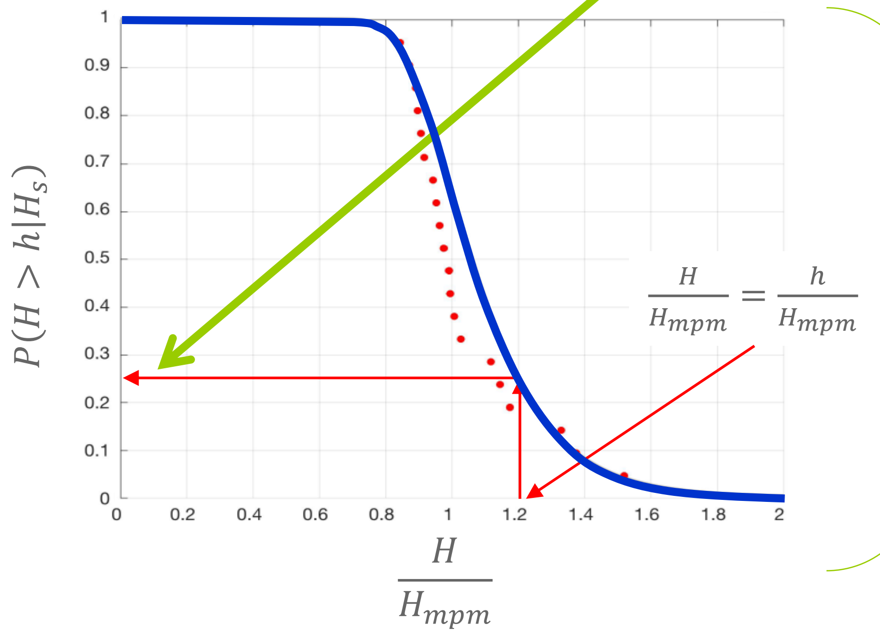
probability density of y conditional on $\theta = b$

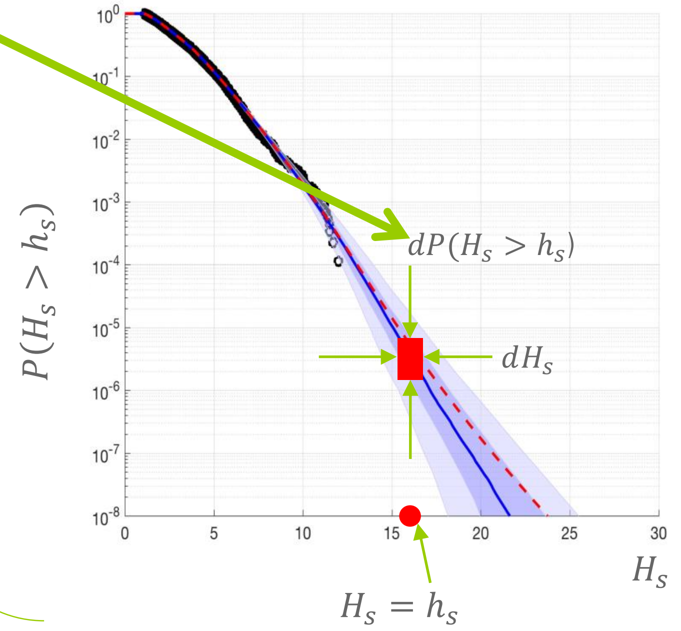
$$p(y|\theta = b) = \frac{p(y,\theta = b)}{\int_0^\infty p(y,\theta = b)dy}$$

divide by marginal to give a valid probability (area =1)

$P(H > h)$ in the long-term – by chain rule (aka convolution in ISO 19901-1)

$$P_{annum}(H > h) = \int_0^\infty P(H > h|H_s) \times p(H_s)dH_s$$





probability of exceedance of largest wave height, $P(H > h)$, in a given 3hr sea state H_s (ie in the short term)

annual probability of the given sea state occurring (ie in the long term)

probability of exceedance for largest in N (random) events

International Association of Oil & Gas Producers

probability of the crest of an individual wave (E_1) not exceeding a given value (η) in a given sea state $H_s = h$ is:

$$P(E_1 \leq \eta | H_s = h)$$

probability of the crest of another individual wave (E_2) not exceeding the same value (η) in $H_s = h$ is:

$$P(E_2 \leq \eta | H_s = h)$$

probability of the larger of crest elevations ($E_1$and E_2) not exceeding a given value (η) in $H_s = h$ (assuming independence ie far apart in the sea state) is:

$$P\Big((E_1 \leq \eta | H_s = h) \text{ and } (E_2 \leq \eta | H_s = h)\Big) = P(E_1 \leq \eta) \times P(E_2 \leq \eta) = \prod_{i=1}^{2} P(E_i \leq \eta) = P\left(\max_{i=1,2}[E_i] \leq \eta\right)$$

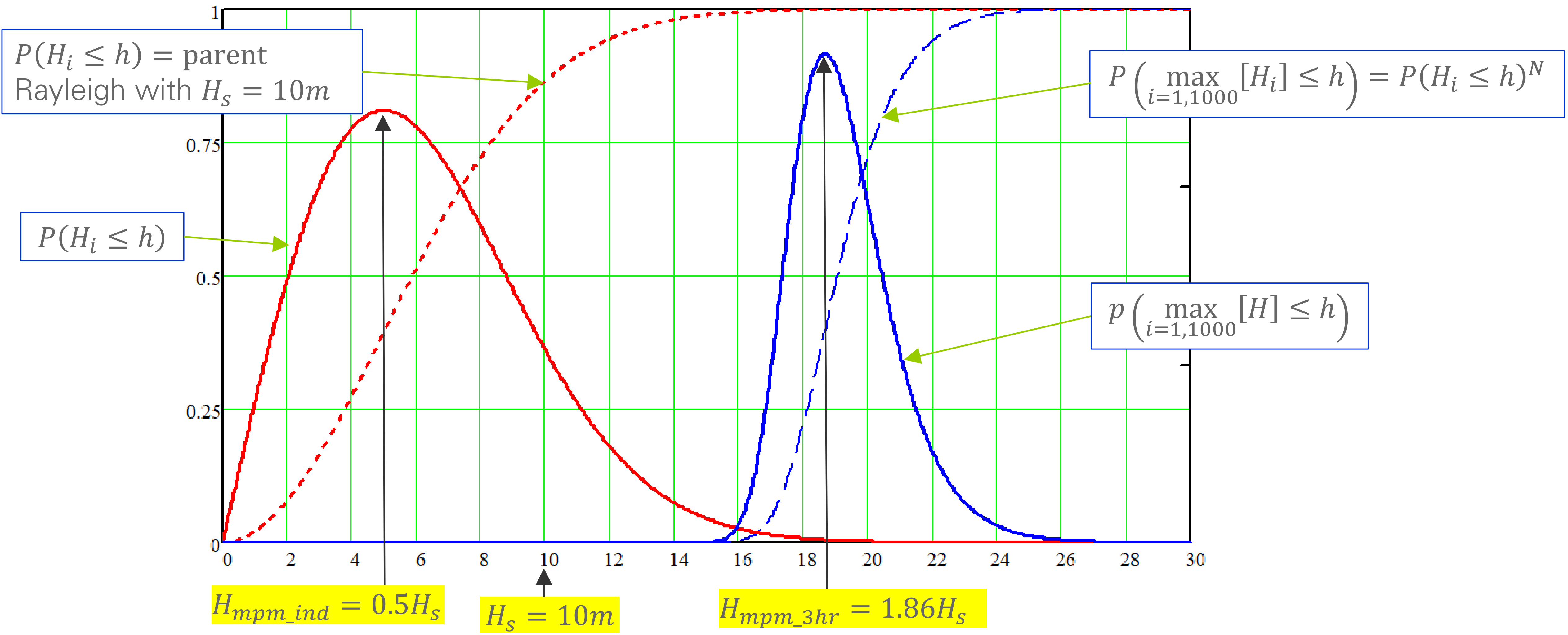
probability of the largest crest elevation in N waves exceeding a given value (η) in $H_s = h$ is

$$P\left(\max_{i=1,N}(E_i) > \eta \middle| H_s = h\right) = 1 - \prod_{i=1}^{N}(E_i \leq \eta | H_s = h)$$

NB probability of the smallest crest elevation in N waves exceeding a given value (η) in $H_s = h$ is $\prod_{i=1}^{N} P(E_i > \eta)$

distributions of individual & largest wave ht for a given sea state with duration 3hr (N=1000)

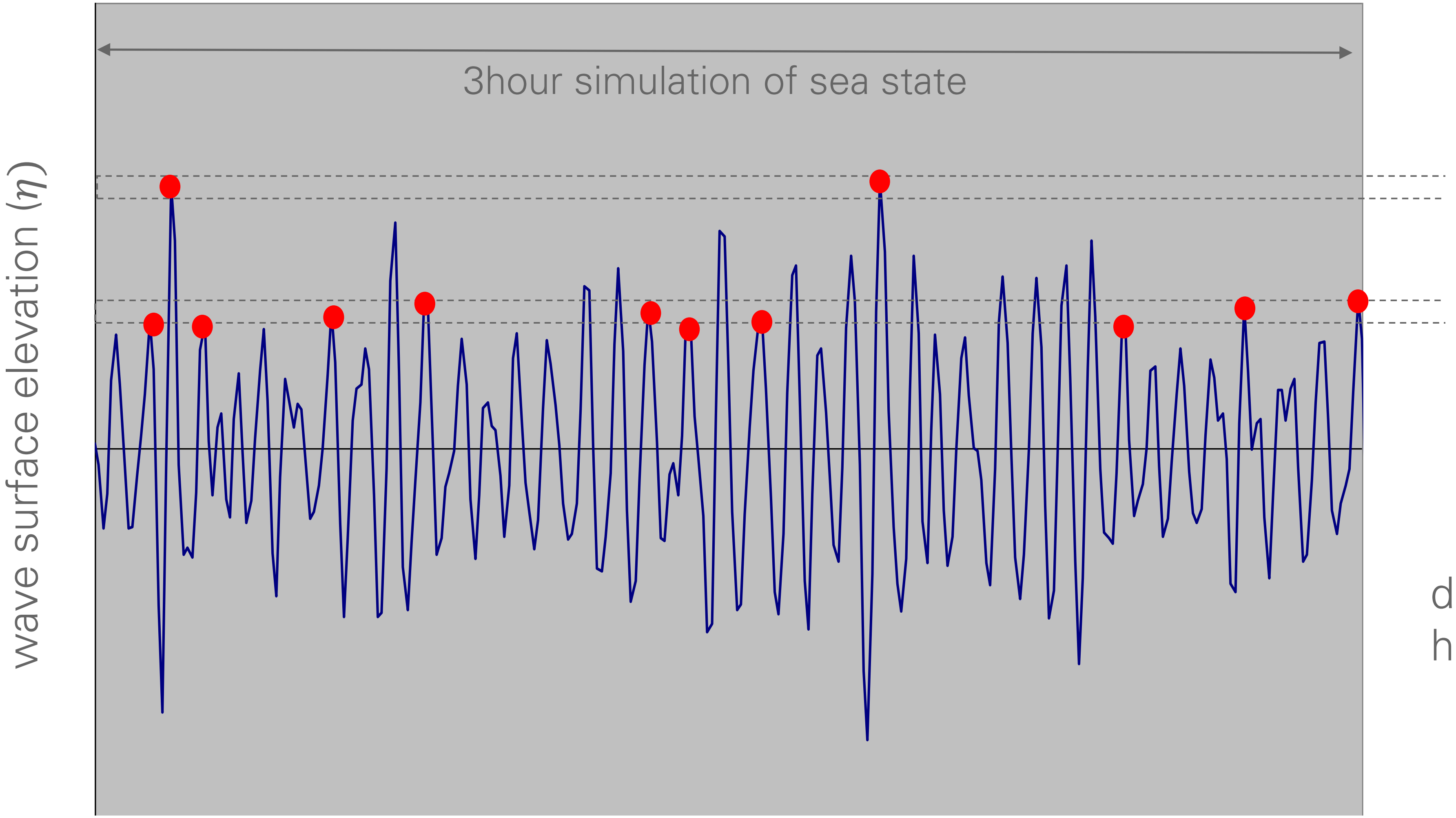
International Association of Oil & Gas Producers



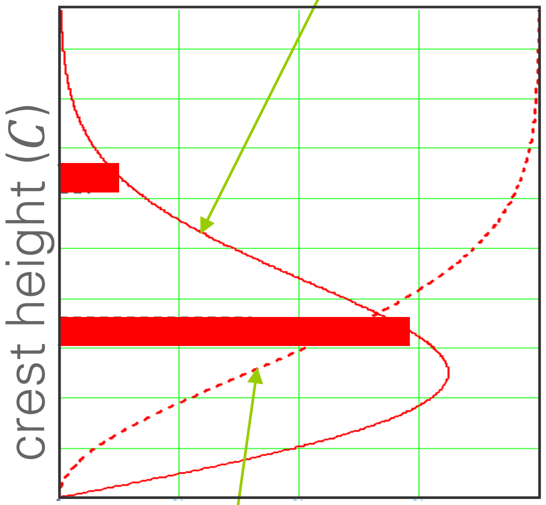
distribution of individual crest ht – given a 3hr sea stat

International Association of Oil & Gas Producers

crest heights shown by red dots



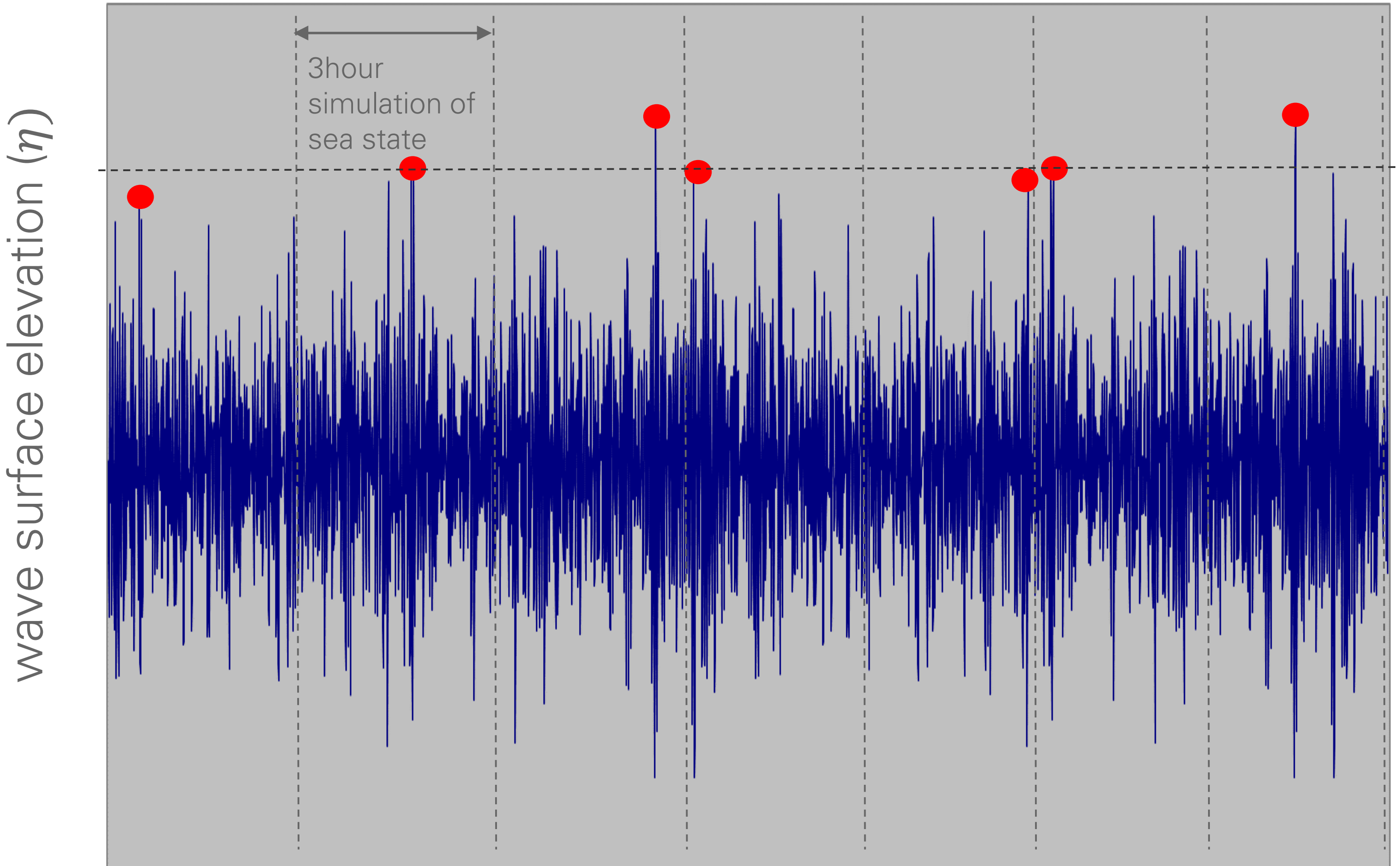
pdf of individual crest height in a given sea state



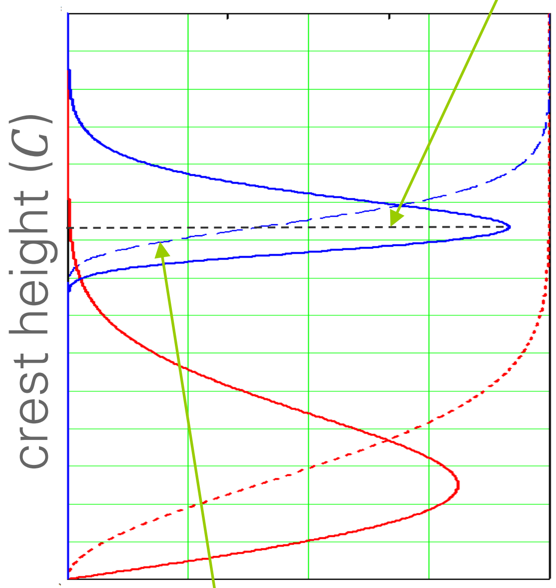
distribution of individual crest height in a given sea state

distribution of largest crest ht – given a 3hr sea state

max crest heights (in 3 hrs) shown by red dots



mpm crest height, C_{mpm}, in a sea state



distribution of largest crest height in a given sea state

statistics of extremes – GEV and GPD

International Association of Oil & Gas Producers

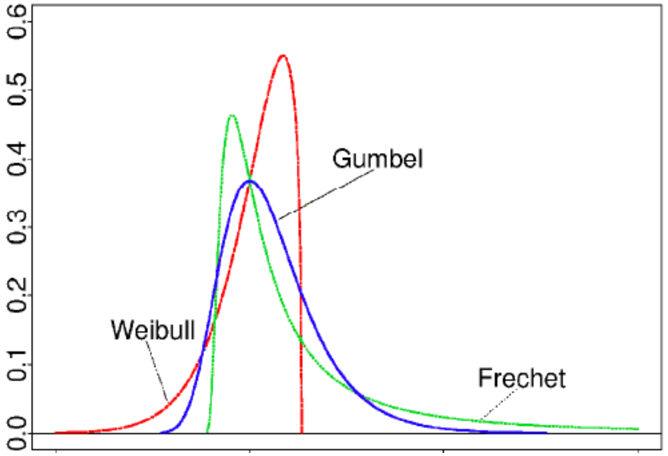
Generalised Extreme Value GEV distribution

$$P(X \leq x) = exp\left(-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right) \text{ if } \xi \neq 0$$

$$P(X \leq x) = exp\left(-exp\left(\frac{x-\mu}{\sigma}\right)\right) \text{ if } \xi = 0$$

μ = location parameter
σ = scale parameter
ξ = shape parameter

GEV is distribution of extreme



Gumbel $\xi = 0$
parent has exponential tail

Pareto (Fréchet) $\xi > 0$
parent has polynomial tail

Weibull $\xi < 0$
parent has upper end point = $[\sigma + \xi(u - \mu)]/|\xi|$

Generalised Pareto distribution GPD

$$P(X - u > x | X > u) = \left[1 + \frac{\xi(x-u)}{\sigma + \xi(u-\mu)}\right]^{-1/\xi}$$

μ = location parameter
σ = scale parameter
ξ = shape parameter
u = threshold

GPD is distribution of data points above a threshold

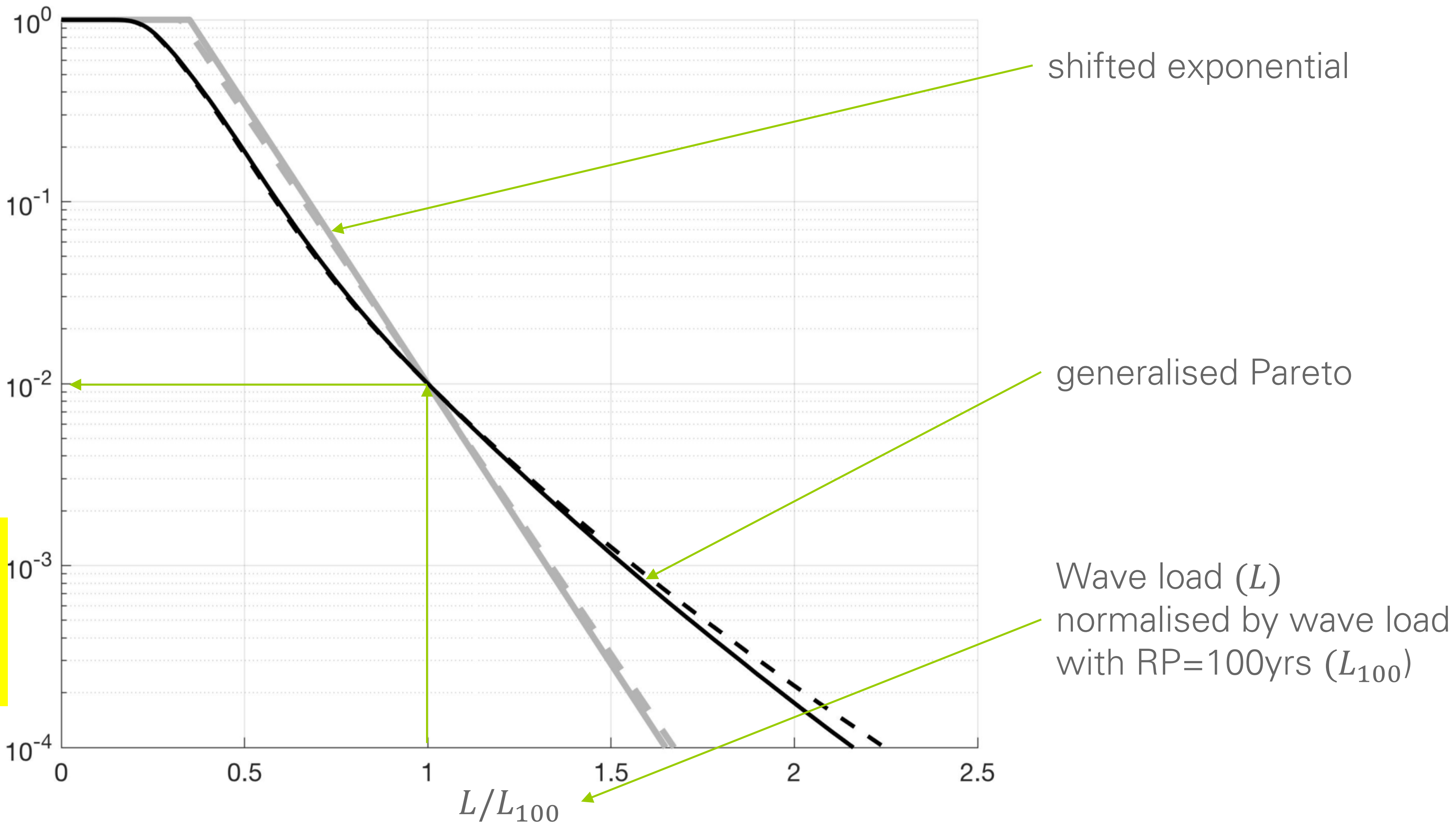
distribution of extreme = (parent distribution)^N tends to GEV asymptotic distribution as N becomes large

form of extreme depends on the form of the tail of the parent distribution

probability of exceedance (extreme values)

annual probability of exceedance

$P_{annum}(L > l|\alpha)$



log scale rather than linear (0 to 1) shows the tail in more detail at extreme values

Poisson probability density function

n is the number of storms, the magnitude of which is greater than a given magnitude m, over a period of length t is Poisson distributed.

$$p(N|\nu_m,t) = \frac{(\nu_m t)^N}{N!} e^{-\nu_m t} \quad \& \quad P(N \le n) = \sum_{i=1}^{n} p(i|\nu_m,t)$$

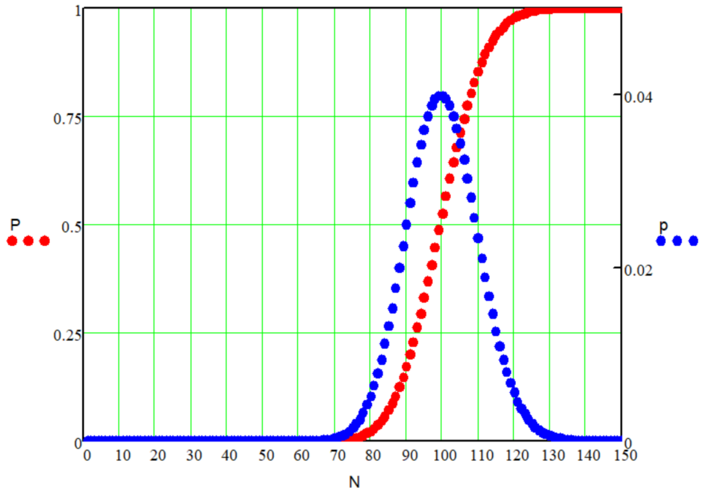
generally, t is taken equal to 1 year, so that ν_m is to be interpreted as the mean annual number of storm occurrences (depends on m) - say ν_m=100

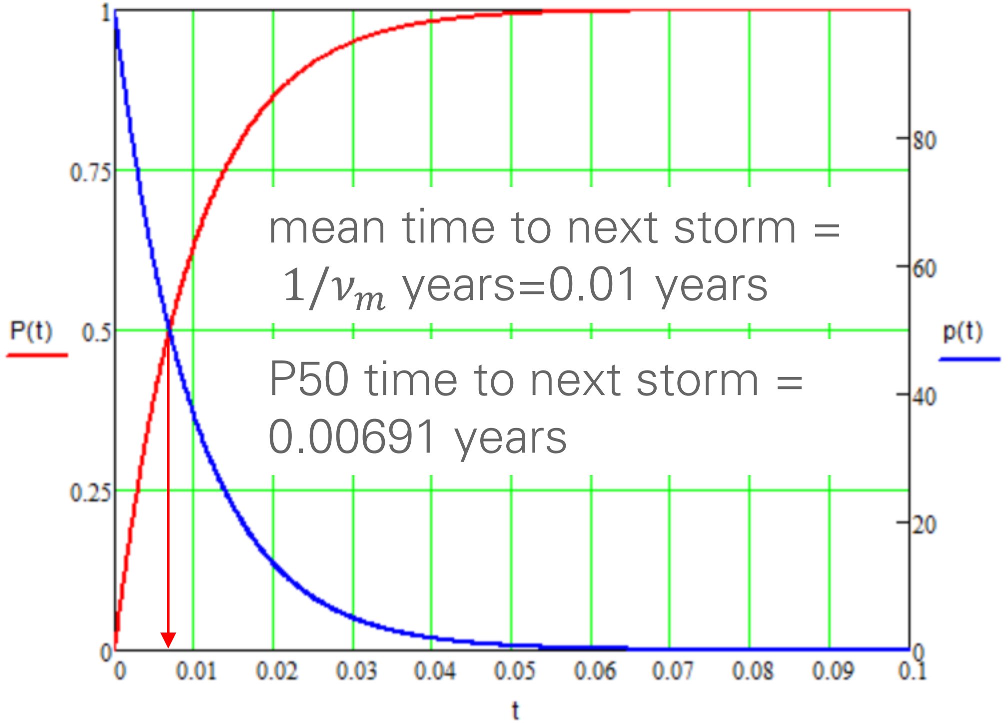
probability that the time taken for the next storm (with magnitude greater than m) to arrive, ie the waiting time T, is less than or equal to t is:

$$P(T \le t) = 1 - \exp(-\nu_m t) \qquad \& \qquad p(T) = \nu_m \exp(-\nu_m T)$$

if time to next storm is t then number of storms during the waiting time T is zero (ie N=0):

$$p(0|\nu_m,t) = \frac{(\nu_m t)^0}{0!} e^{-\nu_m t} = e^{-\nu_m t} = 1 - P(T \le t)$$



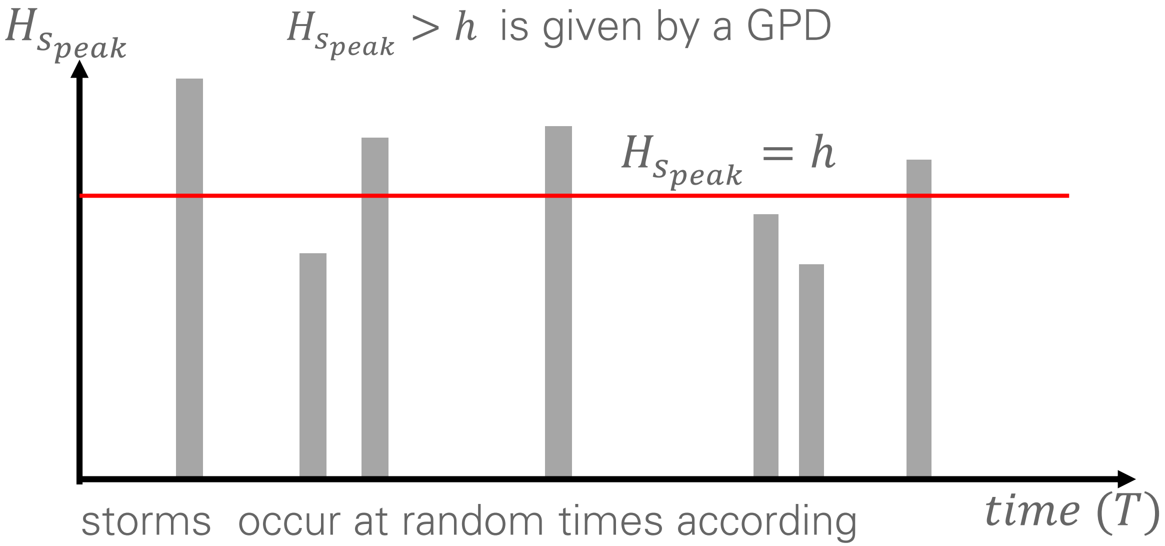


Poisson spike process

– used to describe time-dependent events (eg wave loading due to discrete but infrequent storms)

– probability that the time to next storm (ie waiting time T is $< t$) has an exponential distribution:

Given a storm event the probability of $H_{s_{peak}} > h$ is given by a GPD



storms occur at random times according to a Poisson process with intensity of ν storms per annum

$$P(T \leq t) = 1 - \exp(-\nu t)$$

where ν is the mean annual number of storm occurrences

mean annual number of storms with $H_{s_{peak}} > h$ is

$$\nu_{H_{s_{peak}} > h} = \nu \times P\left(H_{s_{peak}} > h \middle| \mathrm{RS}\right)$$

probability of storms with $H_{s_{peak}} \geq h$ arriving per year is

$$P\left(T \leq 1\ \mathrm{year} \middle| H_{s_{peak}} > h\right) = 1 - \exp\left[-\nu_{H_{s_{peak}} > h} \times 1\ \mathrm{year}\right]$$

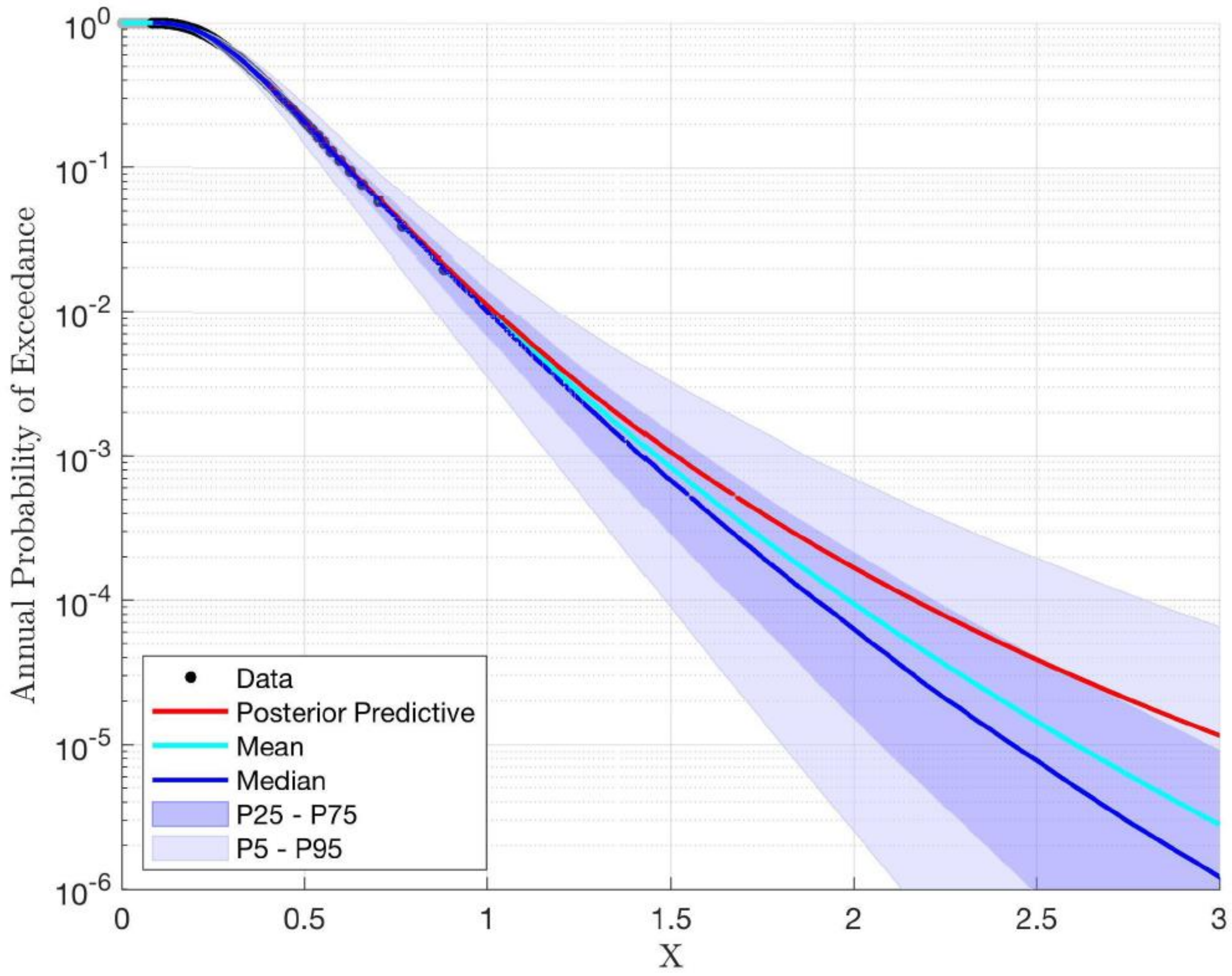
$$P_{annual}\left(H_{s_{peak}} > h\right) = 1 - \exp\left[\nu P\left(H_{s_{peak}} > h \middle| \mathrm{RS}\right)\right] \cong \nu P\left(H_{s_{peak}} > h \middle| \mathrm{RS}\right) \text{ for small } \nu P\left(H_{s_{peak}} > h \middle| \mathrm{RS}\right)$$

credible interval (for Bayesian inference)

Quantiles
are points in a distribution that relate to the rank order of values in that distribution.

Percentiles
are descriptions of quantiles relative to 100; so the 75th percentile (upper quartile) is 75% or three quarters of the way up an ascending list of sorted values of a sample.

Credible interval
Credible interval is a "Bayesian confidence interval", but unlike frequentist confidence intervals, credible intervals have a very intuitive interpretation: the 90% credible interval contains the true parameter value ($\boldsymbol{\theta}$) with 90% probability.



Bayesian inference (1)

International Association of Oil & Gas Producers

Bayesian inference key points…

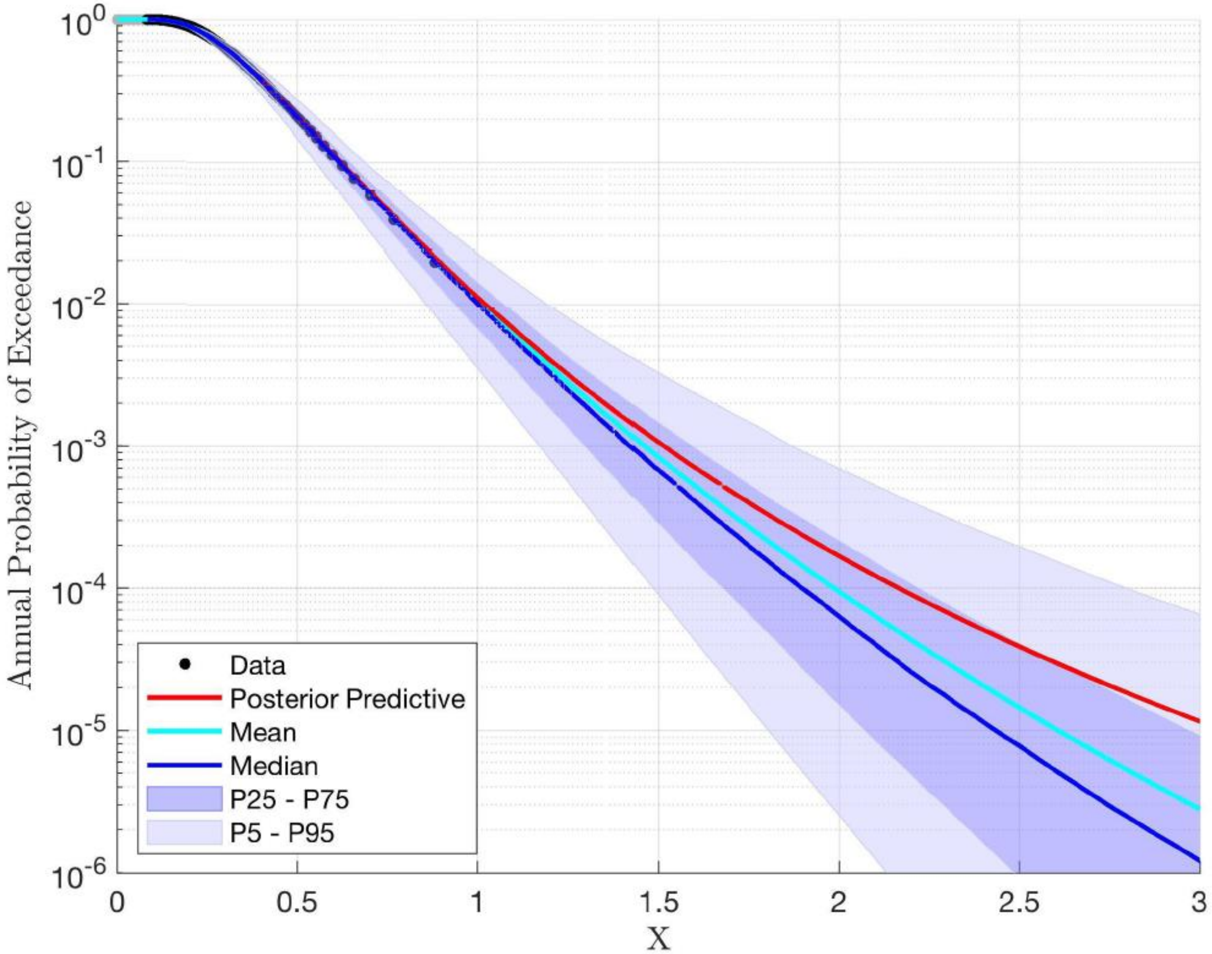
1) uses prior knowledge of parameter distribution

   ie prior distribution of parameters

2) uses available data together with the prior

   ie posterior distribution of parameters

3) gives the uncertainty explicitly



Bayesian inference (2)

International Association of Oil & Gas Producers

$$P(\theta|y) = \frac{P(y|\theta) \times P(\theta)}{P(y)}$$

$$posterior = p(\boldsymbol{\theta}|\boldsymbol{h}_{sp\ data}) = \frac{p(\boldsymbol{h}_{sp\ data}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})}{p(\boldsymbol{h}_{sp\ data})} = \frac{p(\boldsymbol{h}_{sp\ data}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})}{C} = \frac{likelihood \times prior}{C}$$

where
$\boldsymbol{\theta} = [\mu, \sigma, \xi]$ is the vector of parameters for the GPD
$\boldsymbol{h}_{sp\ data}$ is a vector of values of peak H_s in each storm in the metocean long term simulation

Bayes rule calculates probability densities for $[\mu, \sigma, \xi]$ given the data of peak H_s in each storm
a "continuous family" of GPD fits is obtained, the full posterior distribution is used in the LOADS method

The calculation is performed by sampling using MCMC. MCMC doesn't need to know the denominator as it samples in proportion to the relative magnitude of the posterior rather than the absolute.
The samples are then normalised to give a valid posterior pdf.

Bayesian inference (3)

$$posterior = p(\boldsymbol{\theta}|\boldsymbol{h}_{sp\ data}) = \frac{p(\boldsymbol{h}_{sp\ data}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})}{p(\boldsymbol{h}_{sp\ data})} = \frac{p(\boldsymbol{h}_{sp\ data}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})}{C} = \frac{\boxed{likelihood} \times prior}{C}$$

$$p(\boldsymbol{\theta}|\boldsymbol{h}_{sp\ data}) = \frac{\prod_{i=1}^{N_{data}} p\left(h_{sp\ data_i}|\boldsymbol{\theta}\right) \times p(\boldsymbol{\theta})}{\int \prod_{i=1}^{N_{data}} p\left(h_{sp\ data_i}|\boldsymbol{\theta}\right) \times p(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{\boxed{\prod_{i=1}^{N_{data}} p\left(h_{sp\ data_i}|\boldsymbol{\theta}\right)} \times p(\boldsymbol{\theta})}{C}$$

$$p\left(\begin{matrix} h_{sp\ data_{i=1}} \\ h_{sp\ data_{i=2}} \end{matrix}\middle|\boldsymbol{\theta}\right) = p\left(h_{sp\ data_{i=1}}|\boldsymbol{\theta}\right) \times p\left(h_{sp\ data_{i=2}}|\boldsymbol{\theta}\right) = \prod_{i=1}^{2} p\left(h_{sp\ data_i}|\boldsymbol{\theta}\right)$$

Bayesian inference (4)

Determine using Bayesian inference with a Generalised Pareto distribution GPD

$$P(H_{sp} - u > h_{sp} | H_{sp} > u) = \left[1 + \frac{\xi(h_{sp} - u)}{\sigma + \xi(u - \mu)}\right]^{-1/\xi}$$

μ = location parameter
σ = scale parameter
ξ = shape parameter
u = threshold
H_{sp} = peak significant wave ht in a storm. Say we have 1200 years of data $h_{sp\ data_i}$ $i = 1, N$

$$P\left(H_{sp} - u > h_{sp\ data_i} \middle| H_{sp} > u, \xi, \sigma, \mu\right) = \left[1 + \frac{\xi\left(h_{sp\ data_i} - u\right)}{\sigma + \xi(u - \mu)}\right]^{-1/\xi}$$

Sampling the above for each $h_{sp\ data_i}$ for a range of parameters ξ_j, σ_j, μ_j and then taking the product over $i = 1, N$ gives the likelihood

$P(H_{sp} > h|\boldsymbol{h}_{sp\ data})$ - posterior predictive prob. exceedance

International Association of Oil & Gas Producers

sampling distribution for future observations of Hs given the GPD parameters

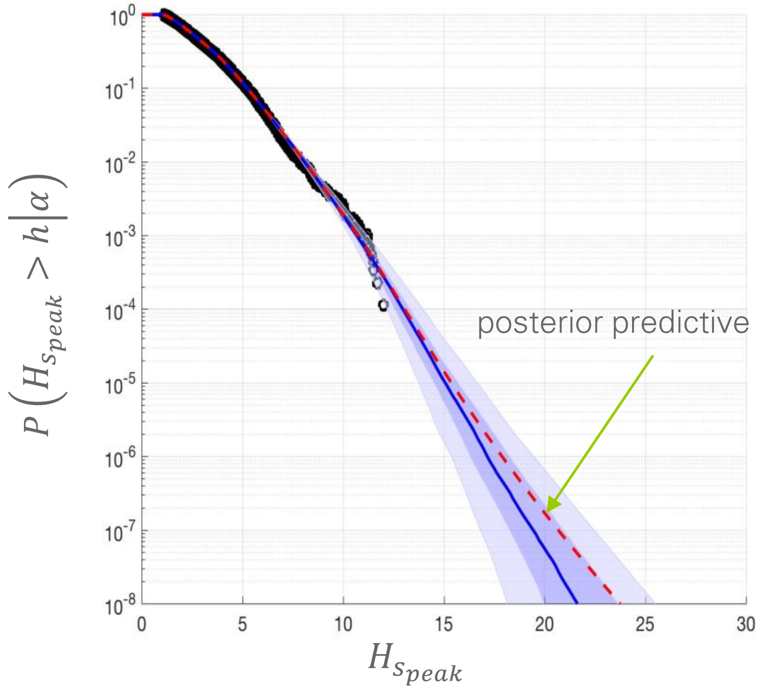
posterior distribution of the parameters given past observations of Hs (ie data)

posterior predictive distribution is a conditional expectation (conditioned on the observed data) weighted by the parameter values from the posterior distribution

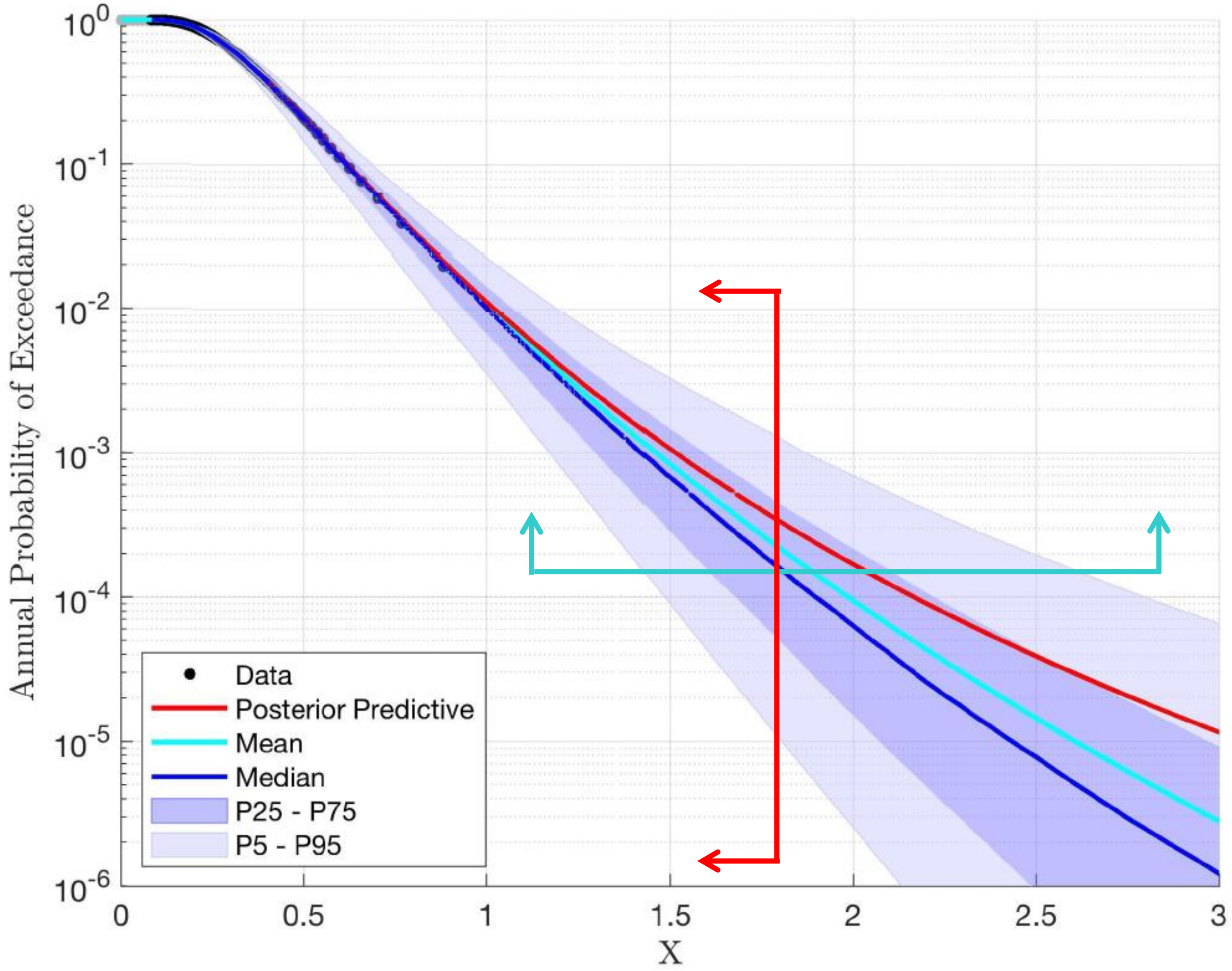
$$\text{posterior predictive PDF} = p(h|\boldsymbol{h}_{sp\ data}) = \int p(h|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}|\boldsymbol{h}_{sp\ data})d\boldsymbol{\theta} = \mathbb{E}\left(p(h|\boldsymbol{\theta})\middle|\boldsymbol{h}_{sp\ data}\right)$$

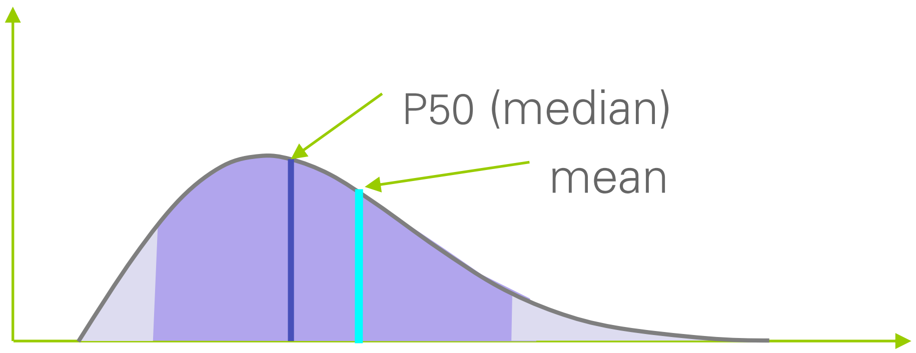
$$\text{posterior predictive CCDF} = P(H_{sp} > h|\boldsymbol{h}_{sp\ data}) = \int_{h=H_{sp}}^{\infty} p(h|\boldsymbol{h}_{sp\ data})dh$$

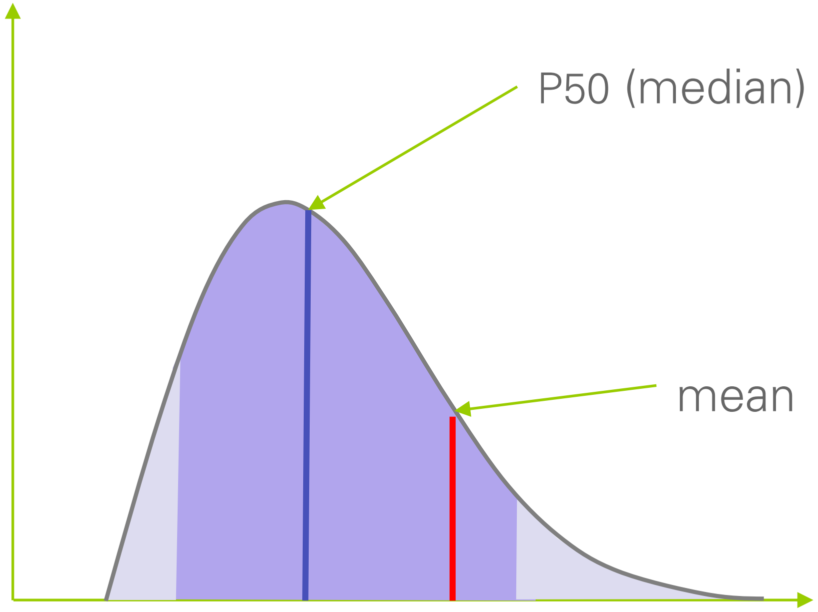
the posterior predictive distribution takes into account the uncertainty of the parameter estimates, which is quantified by the posterior distribution.



mean of probabilities v mean of values







posterior predictive
prob. of exceedance

Earthquake Spectra, Volume 21, No. 3,
pages 879–886, August 2005;
Earthquake Engineering Research Institute

The Case for Using Mean Seismic Hazard

Robin K. McGuire,[a) M.EERI, C. Allin Cornell,[b)
M.EERI, and Gabriel R. Toro,[a) M.EERI

International Association of Oil & Gas Producers

Complete probabilistic seismic hazard analyses incorporate epistemic uncertainties in assumptions, models, and parameters, and lead to a distribution of annual frequency of exceedance versus ground motion amplitude (the "seismic hazard"). For decision making, if a single representation of the seismic hazard is required, it is always preferable to use the mean of this distribution, rather than some other representation, such as a particular fractile. Use of the mean is consistent with modern interpretations of probability and with precedents of safety goals and cost-benefit analysis.
[DOI: 10.1193/1.1985447]

INTRODUCTION

Estimates of earthquake ground motion hazard involve substantial epistemic uncertainty in the mean frequency of exceedance for a given ground motion or, alternatively, in the ground motion for a given mean frequency of exceedance. We believe that the mean estimate of the mean frequency of exceedance should be the standard when a single estimate is necessary. (Please refer to the Addendum for a clarification of the often misused or misunderstood definitions regarding "hazard," "frequency," and "mean." In the context of that addendum we shall adopt the common shorthand convention of "hazard" for the "frequency of exceedance" and "mean hazard" or "mean frequency of exceedance" for the "mean estimate of the frequency of exceedance.")

This epistemic uncertainty has been known and quantified in the United States since the 1970s (see, for example, McGuire 1977). In seismic hazard studies it is preferable to report the complete epistemic distribution of hazard, because this allows any effects of that uncertainty on risk mitigation decisions to be handled in an explicit, quantitative way. This reporting usually takes the form of presenting four or more hazard curves, say, three fractiles (e.g., 0.10, 0.50—or median—and 0.90) plus the mean hazard curve. This reporting position is supported by a finding of the National Research Council Panel on Seismic Hazard Analysis: "Knowledge of earthquake processes and effects in much of the United States is meager, resulting in considerable uncertainty in seismic hazard estimates. No single measure of the seismic hazard (e.g., a mean or median [estimate]) is adequate to represent this basic lack of understanding; therefore, measures of uncertainty must be transmitted as part of a PSHA [probabilistic seismic hazard analysis]." (NRC 1988) (words in brackets added for clarity).