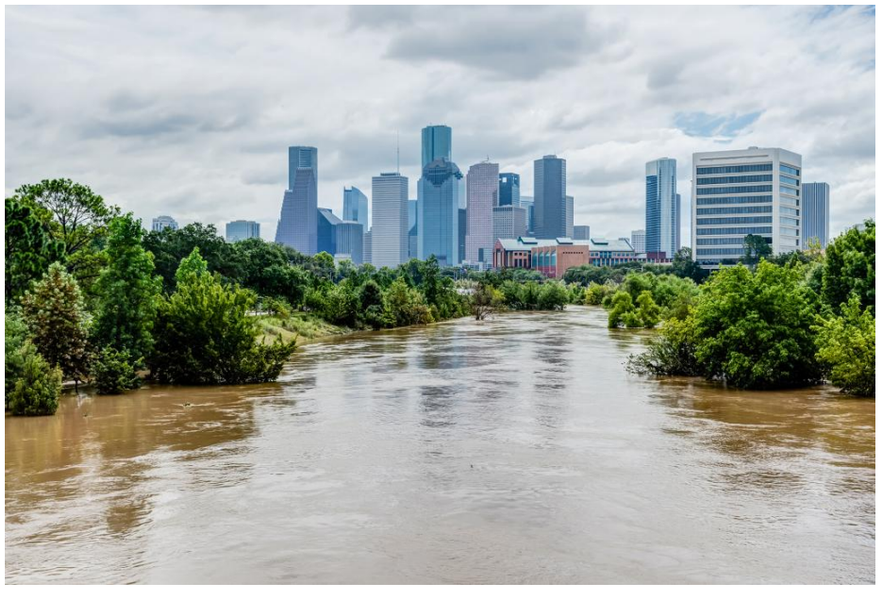
ACM Winter School 2025

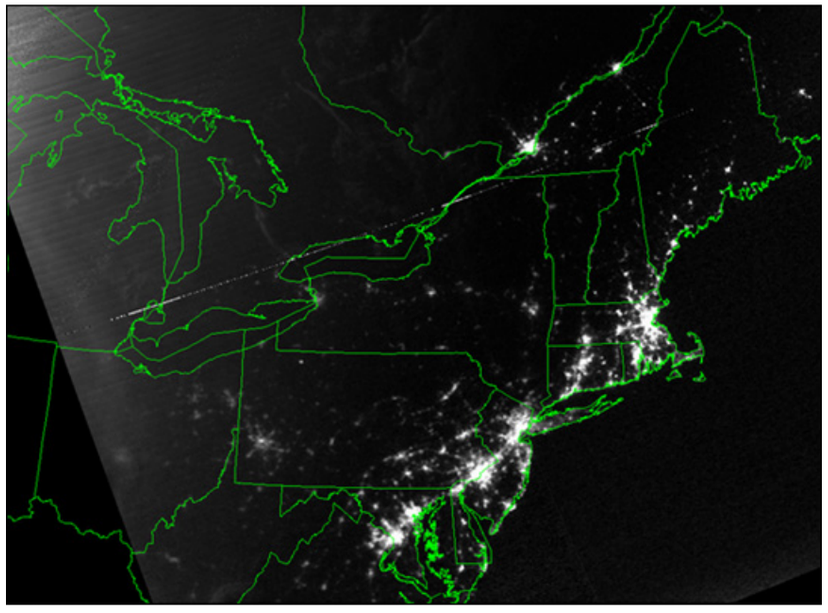
Lecture 1: Input Modelling

© Anand Deo (Indian Institute of Management, Bangalore)

Motivation: decision making under uncertainty



Hurricane Harvey, 2017



Northeast US Blackout, 2003

Indigo Crisis, 2025

Suez Canal Obstruction, 2021

<100 large power outages/year in USA

<1000 large floods globally since 1985

<5 severe airlines crises since 2000

Tens of billions of dollars loss

How can we mitigate risk of extreme failures?

An example: uncertainty and financial portfolios

(What are these sessions all about?)

Consider a portfolio holding a number of liquid assets, for instance investments in d financial sectors. Let $\xi \in \mathbb{R}^d$ be the random vector of one-day losses. Let $\boldsymbol{x} = (x_1, \ldots x_d)$ denote investments in each sector. Then the one-day portfolio loss is $L(\boldsymbol{x}) = \boldsymbol{x}^\intercal \xi$. A risk manager may ask the following questions:

i) How can we build a probabilistic model for one-day losses from historical data, and then generate simulated loss scenarios from this mode?

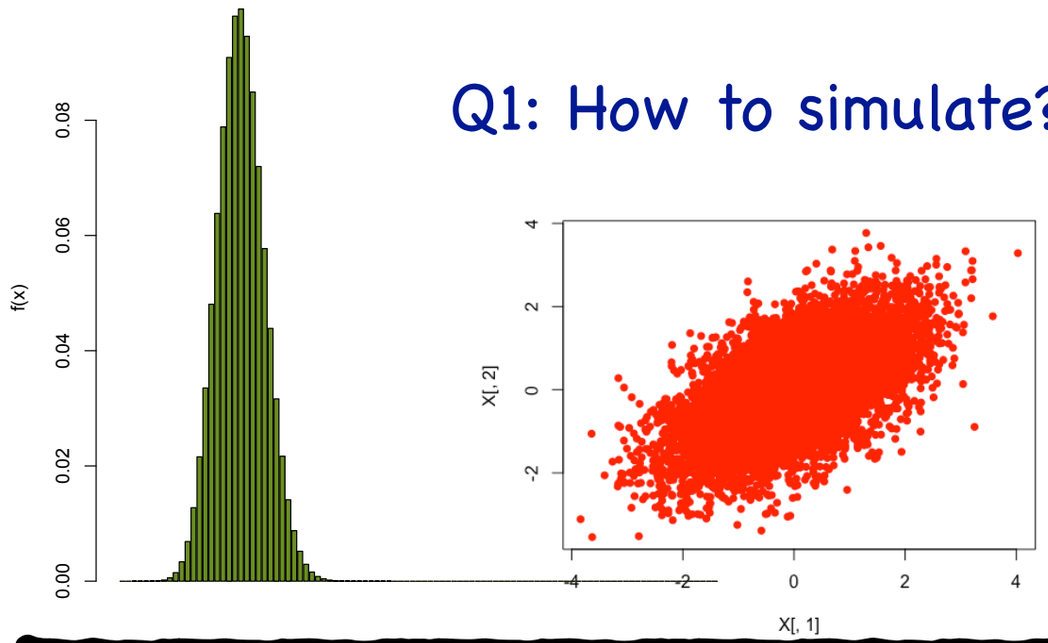
ii) How much capital should be kept aside, so that the probability that the one-day loss exceeds this buffer stays below a prescribed threshold?

iii) How sensitive are the answers to the above questions to assumptions about the tail of the loss distribution, especially for extreme losses?

iv) How can one reduce the sensitivity of risk estimates to uncertainty about the data distribution?

An example: uncertainty and financial portfolios

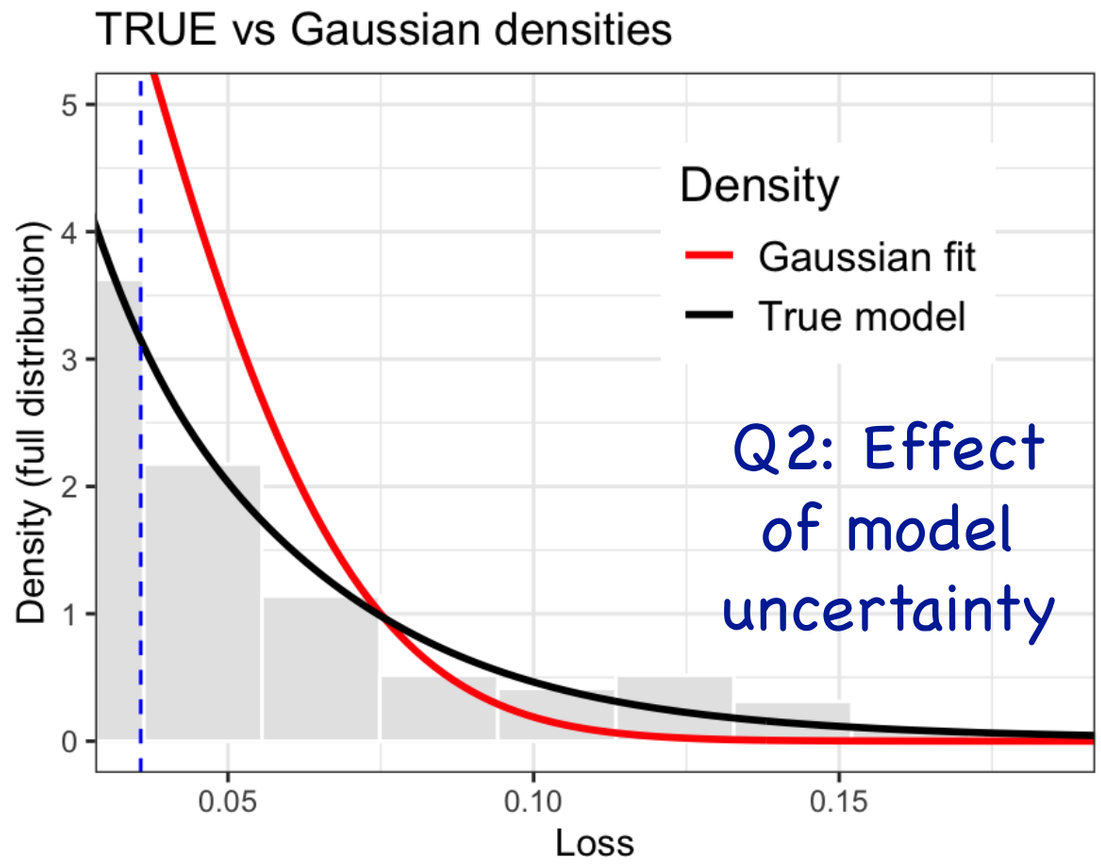
(What are these sessions all about?)



Q1: How to simulate?

Lecture 1: Input Modelling

- Model for data → generation of samples
- Impact of model ambiguity, and potential recourses



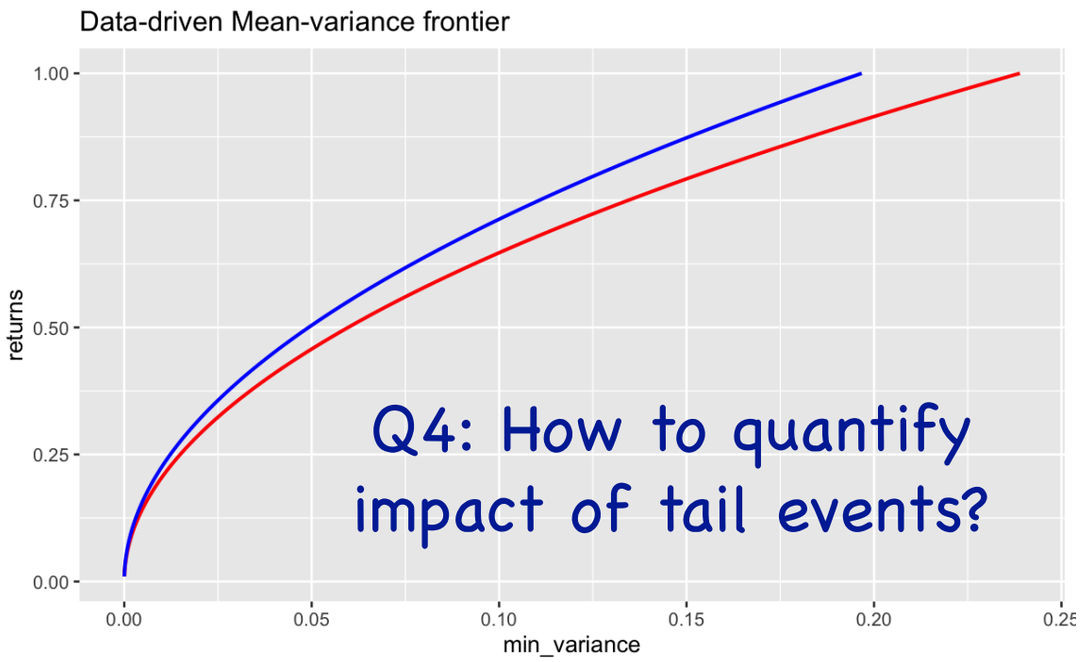
Q2: Effect of model uncertainty

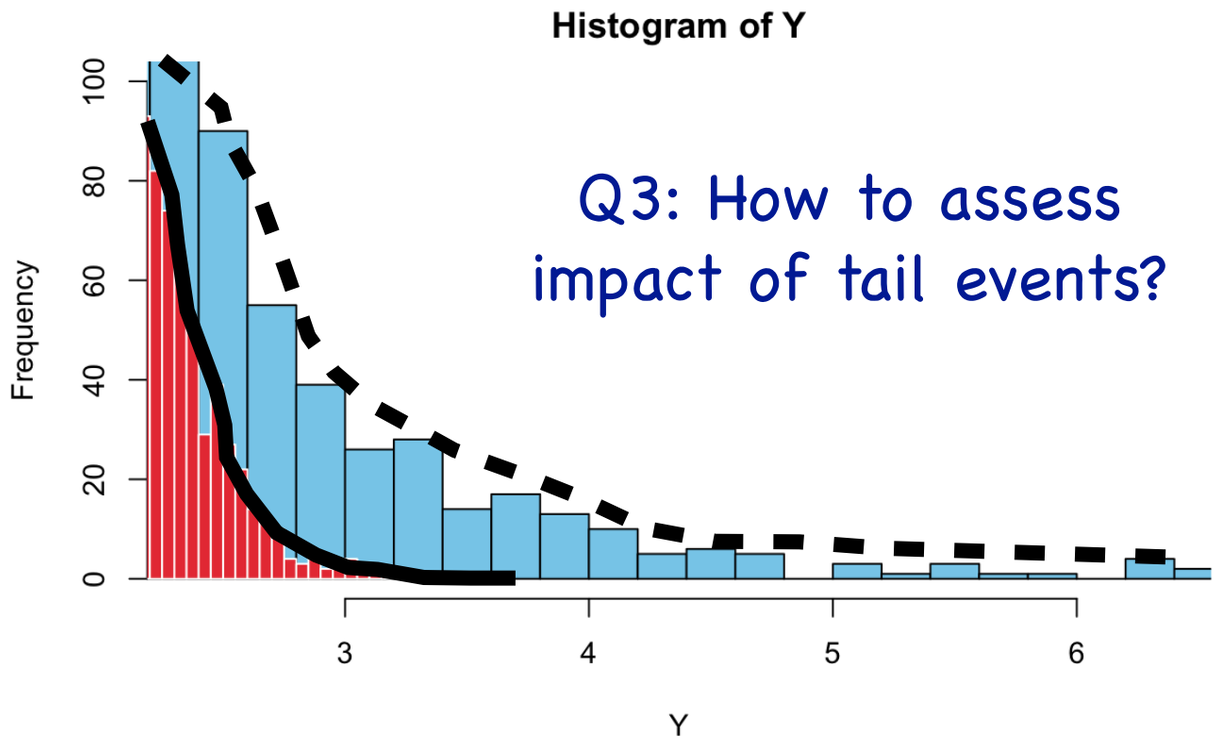
Lecture 2: Copula and Model Risk

- Generating sophisticated input models
- Wrong Model data → poor risk assessment

An example: uncertainty and financial portfolios

(What are these sessions all about?)





Lecture 4: Risk Measures

- How to characterise the impact of uncertainty
- From data —> decisions

Lecture 3: Extreme Value Theory

- Overcoming the model risk in estimation of tails
- Basic principles behind EV design

Why simulate?

(When do direct computations not work?)

Let $\xi \in \mathbb{R}^d$ be the random vector of one-day losses and $x = (x_1, \ldots x_d)$ denote investments in each sector. Task: find how much capital should be kept aside, so that the probability that the one-day loss exceeds this buffer stays below 1%

- Mathematically written, the above task is to find u s.t. $P(x^\intercal \xi > u) = 0.01$.
- For this, one needs to evaluate the integral $\int_{x^\intercal z > u} f_\xi(z) dz$.
- Difficulty: Evaluating multidimensional integrals is HARD!
- Recourse: Generate $\{\xi_1, \ldots, \xi_n\} \sim \xi$ and output an estimate $\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x^\intercal \xi_i > u)$

Guiding principle: to estimate a risk functional $T(P)$, (mean, tail probabilities, quantiles), replace $P \to$ sample (simulation) estimate

Agenda for the course

Session 1

Input Modelling 🚂

The bootstrap

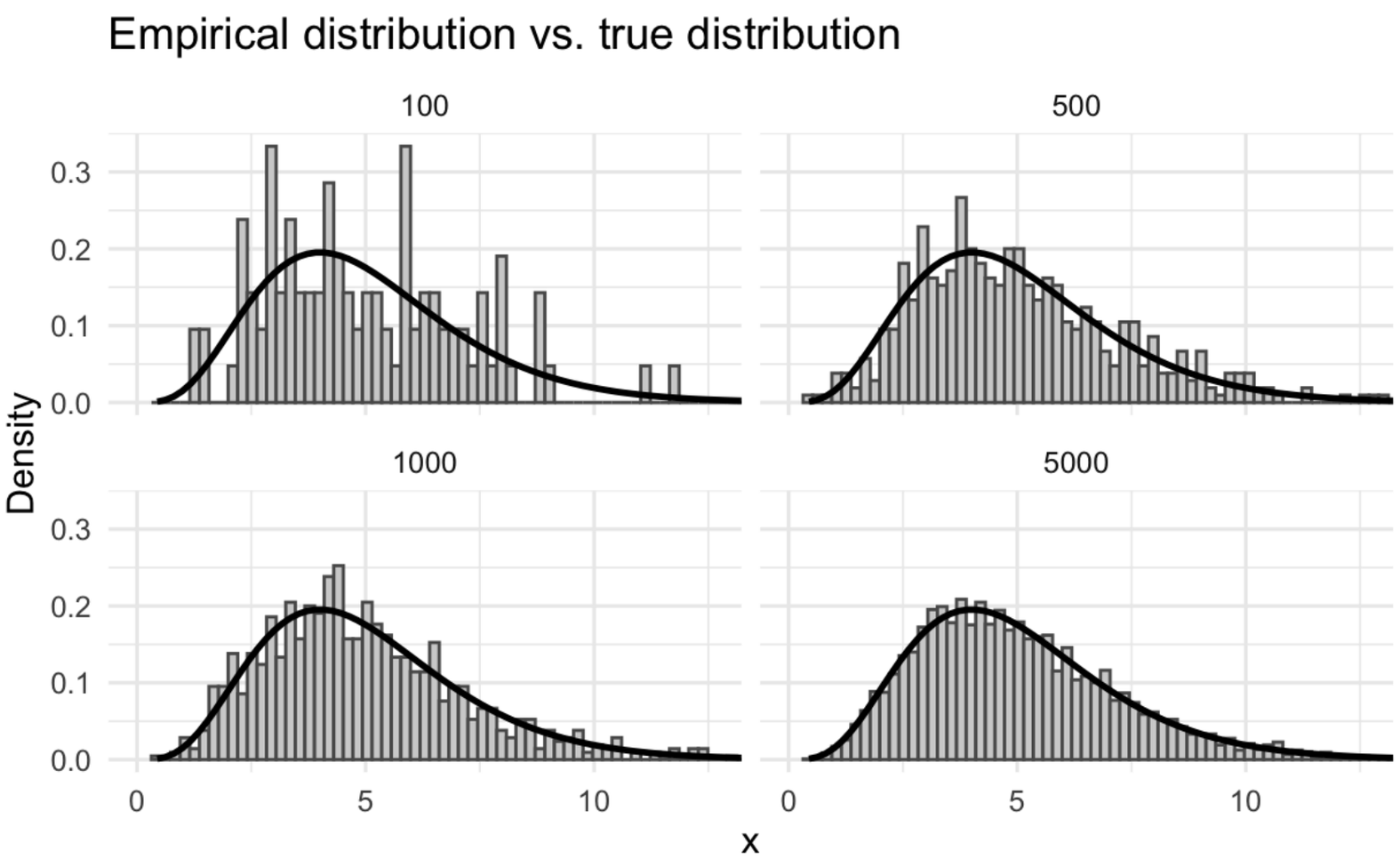
The bootstrap: generating samples directly from data

(A non-parametric approach to simulation)

Suppose we are given historical losses for individual asset classes, without making further parametric assumptions about the data distribution.

IT	FMCG	Healthcare	Metals
-0.046940523	-0.0104490105	0.199355620	0.050143365
0.065850797	0.1776190385	0.058167522	-0.130140429
-0.064688778	-0.0387441708	0.125805716	0.065138189
0.040083871	0.0217020518	-0.001401286	0.135012944

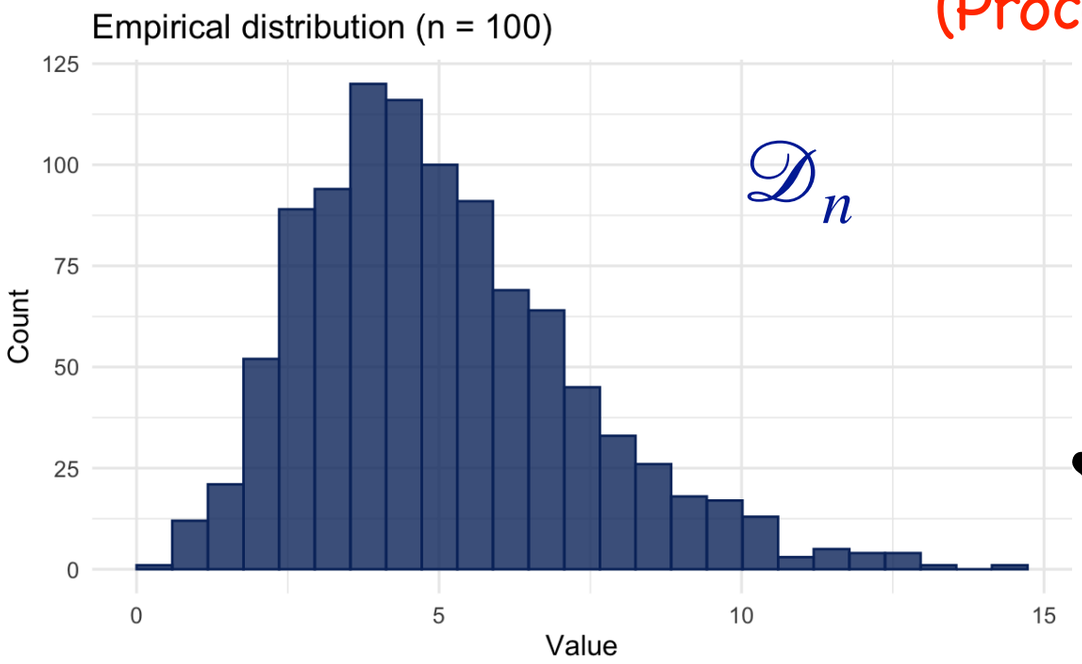
Bootstrap → nonparametric, data-driven input modelling approach



- Empirical distribution P_n → histogram of samples
- It appears that as $n \to \infty$, $P_n \approx P$
- Bootstrap → assume that the model is P_n and generate samples!

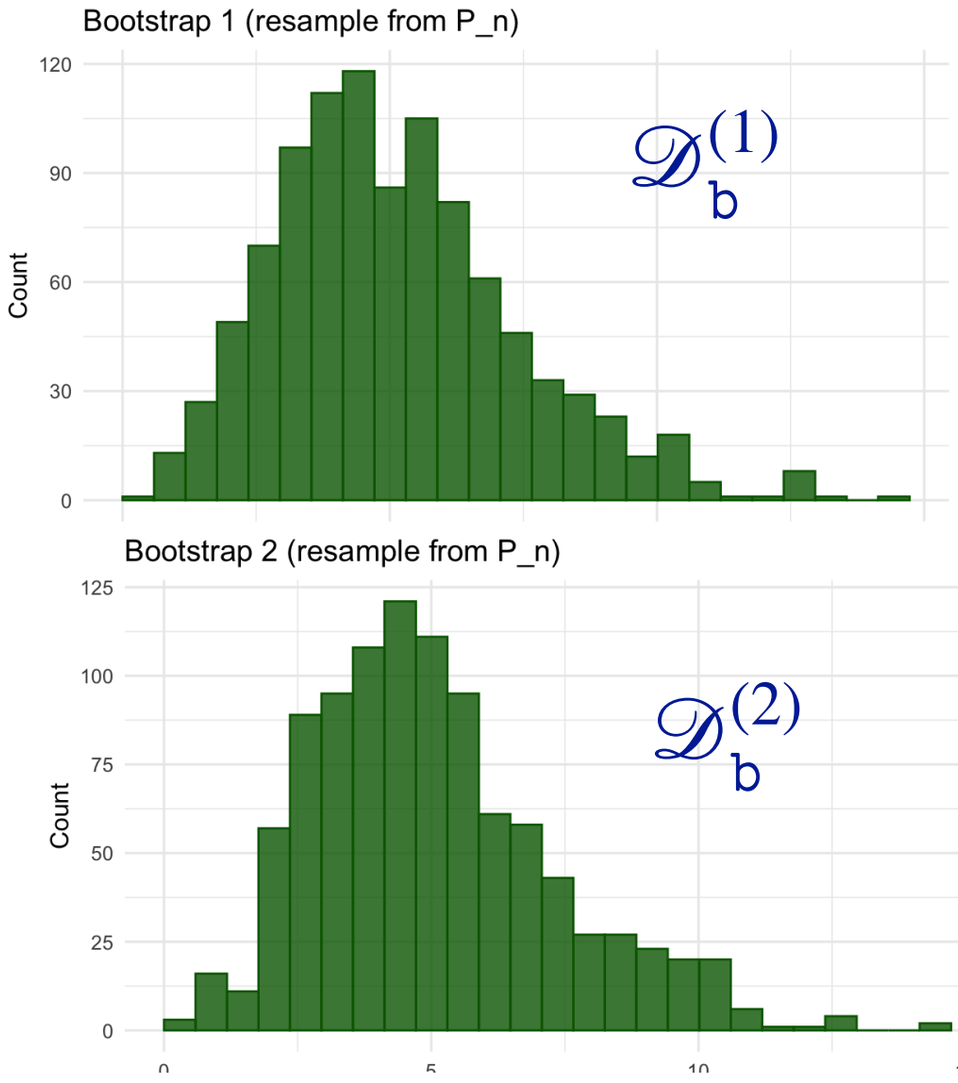
The bootstrap: generating samples directly from data

(Procedure for bootstrap)



$\mathscr{D}_n$

Generate replications



$\mathscr{D}_b^{(1)}$

$\mathscr{D}_b^{(2)}$

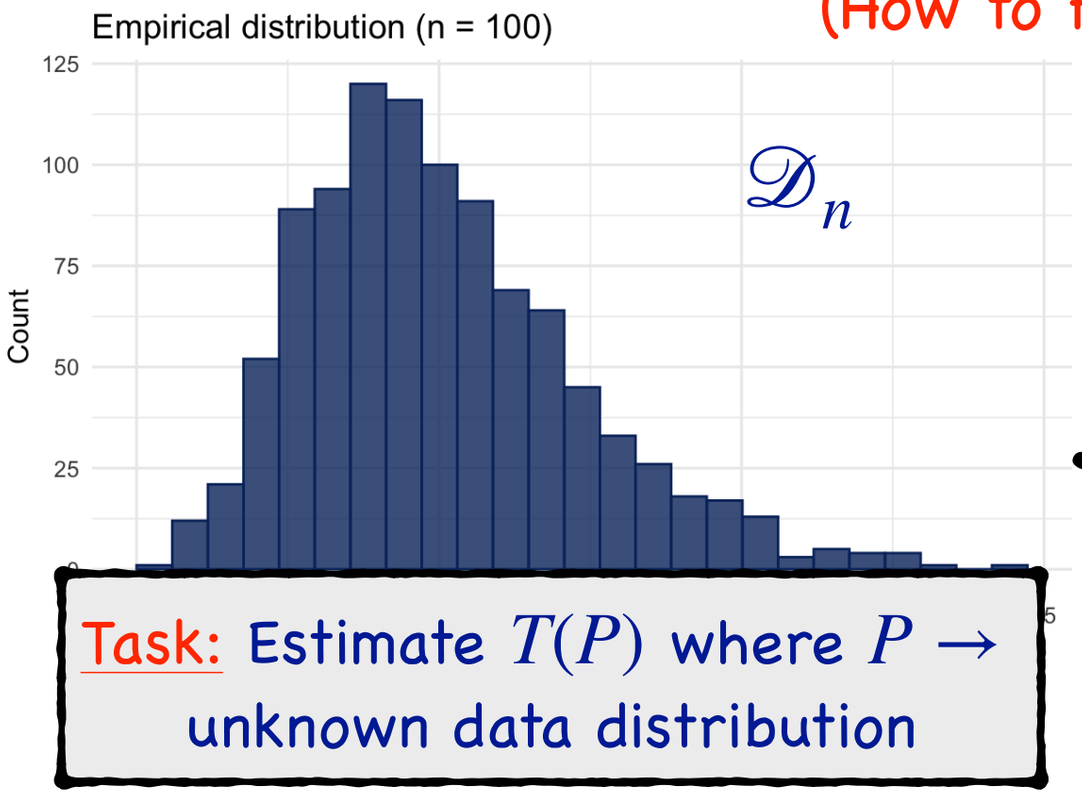
$\mathscr{D}_b^{(M)}$

Procedure: Generation of a bootstrap sample

- Given data $\mathscr{D}_n = (X_1, \ldots, X_n)$ from the distribution of X
- Draw samples $(X_1^\star, \ldots, X_n^\star)$ with replacement from $\mathscr{D}_n$.

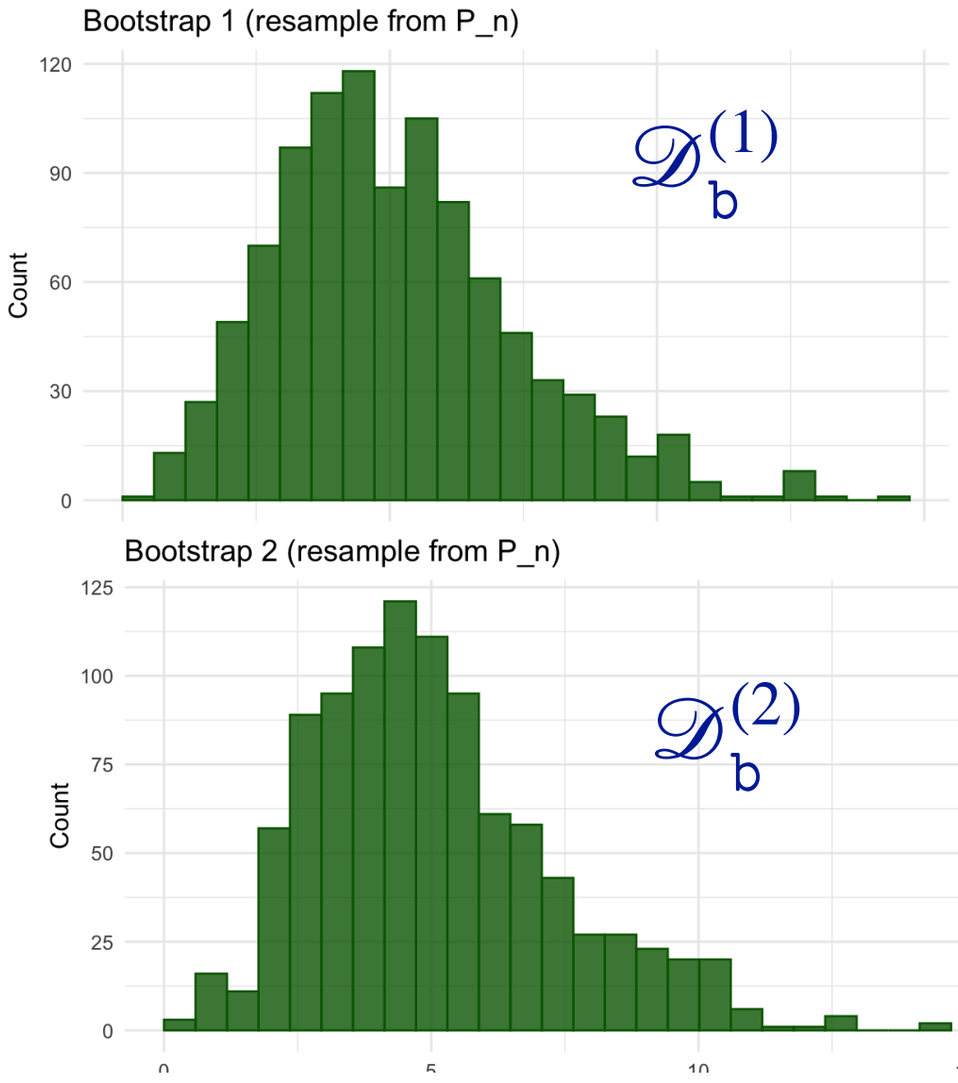
The Bootstrap: Application to parameter estimation

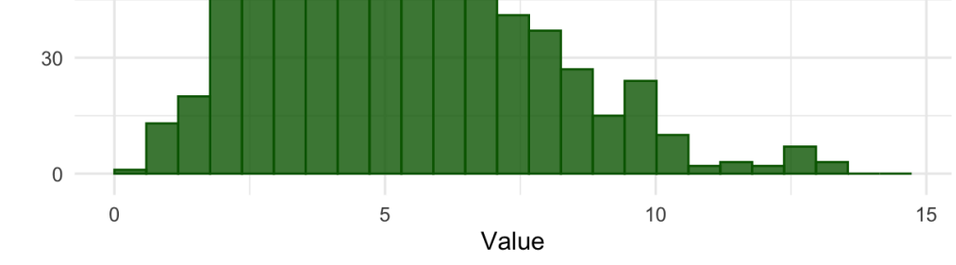
(How to find CI's using bootstrap)



Task: Estimate $T(P)$ where $P \rightarrow$ unknown data distribution

Generate replications





Examples of $T(P)$

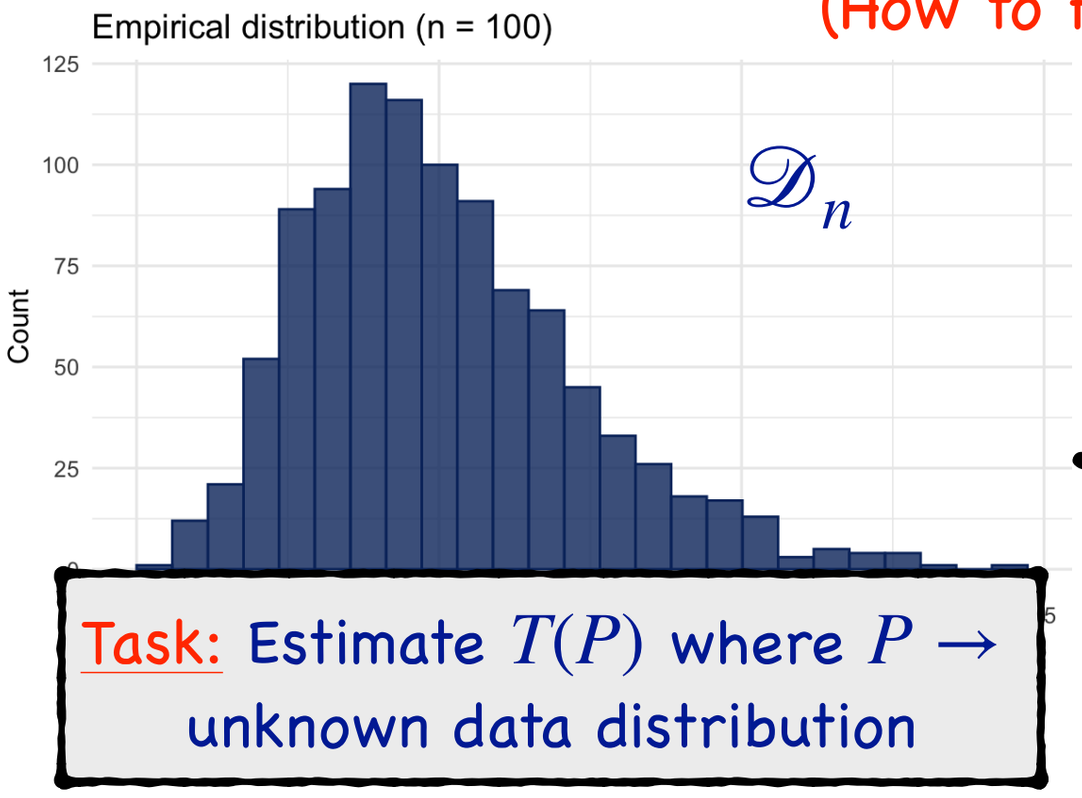
i) $T(P) = E_P[Z] = \int Z dP$ (avg. risk)

ii) $T(P) = \text{Var}(P) = E_P[Z - E_P[Z]]^2$ (variance)

iii) $T(P) = v_{1-q}(P)$ (value at risk)

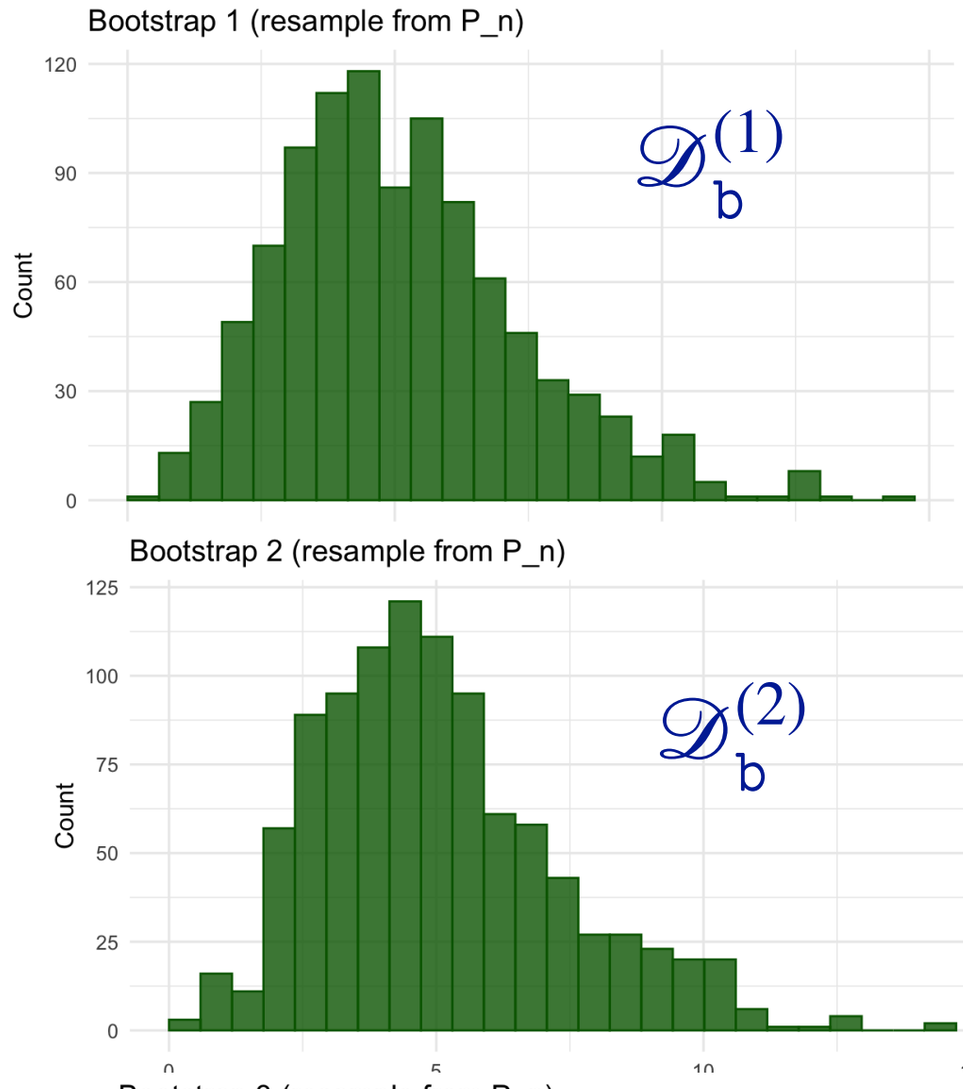
The Bootstrap: Application to parameter estimation

(How to find CI's using bootstrap)



Task: Estimate $T(P)$ where $P \rightarrow$ unknown data distribution

Generate replications

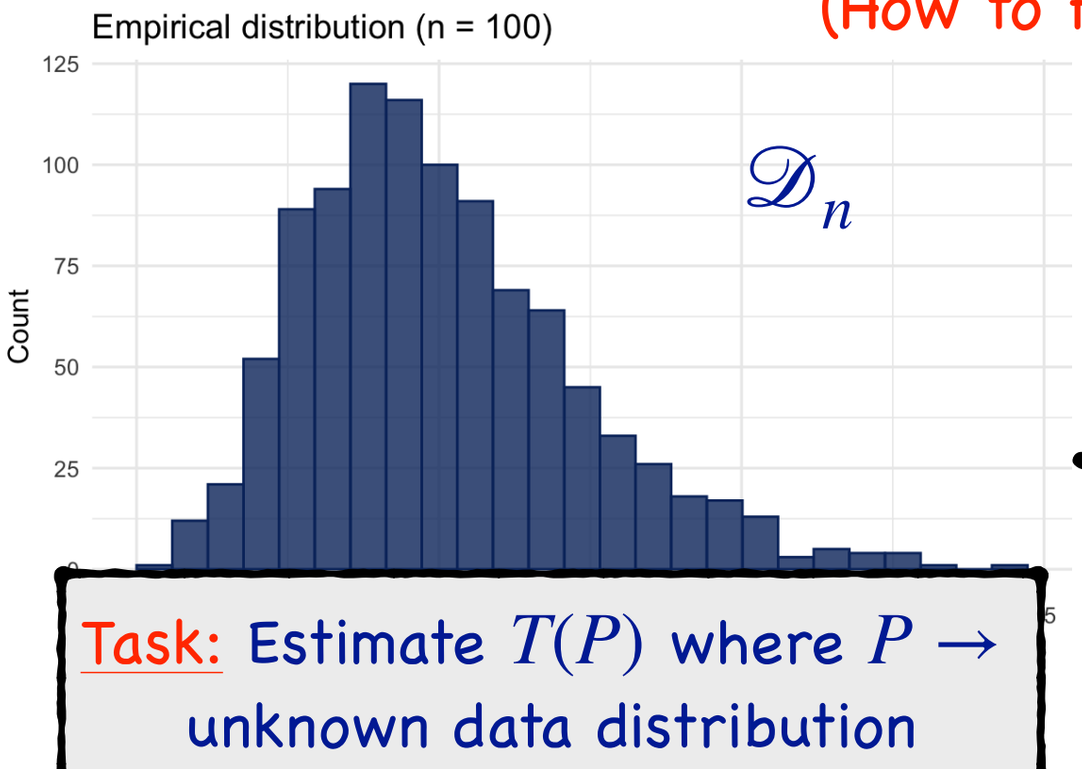


$T^{(1)} = T(P_n^{(i)})$ where $P_n^{(i)} \rightarrow$ empirical distribution of ith bootstrap

Get estimates: $\{T^{(1)}, \ldots, T^{(M)}\}$

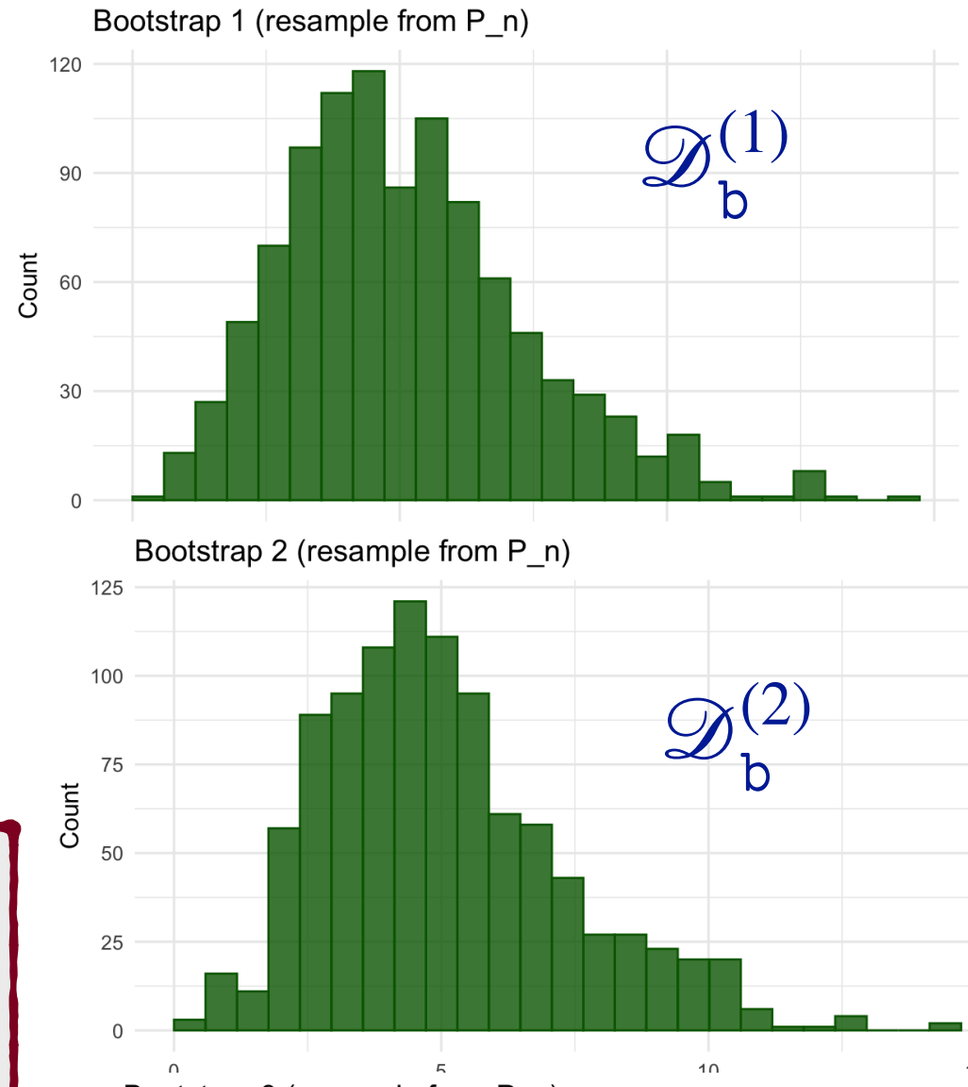
The Bootstrap: Application to parameter estimation

(How to find CI's using bootstrap)



Task: Estimate $T(P)$ where $P \rightarrow$ unknown data distribution

Generate replications



- No Gaussian assumption on sampling distribution!
- Method is purely data-driven!

Return CI:

$$\text{CI} = (q_{\alpha/2}, q_{1-\alpha/2})$$

Get quantiles: $q_p \rightarrow$ pth quantile of estimates

Get estimates: $\{T^{(1)}, \ldots, T^{(M)}\}$

The bootstrap: generating samples directly from data

(Accuracy guarantees for empirical distribution as a model)

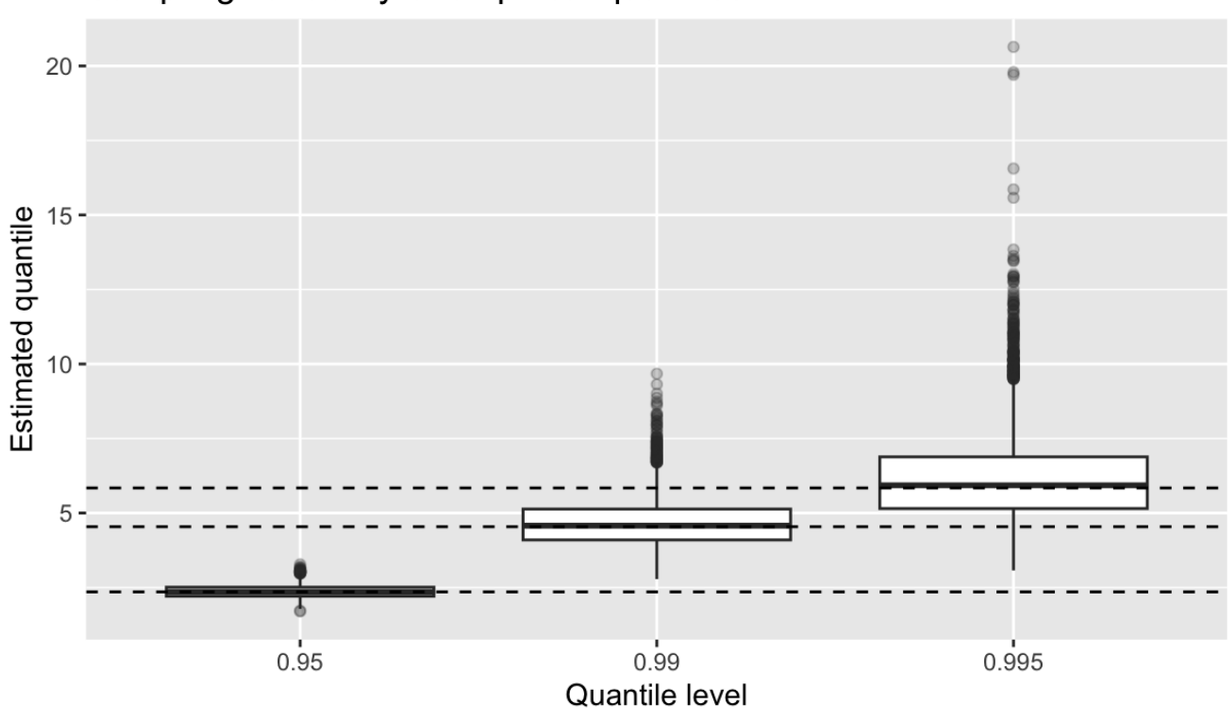
- Bootstrap → Empirical measure P_n is the distribution used to simulate

Result: DKW inequality - if $F_n \rightarrow$ cdf of P_n and $F \rightarrow$ cdf of P then

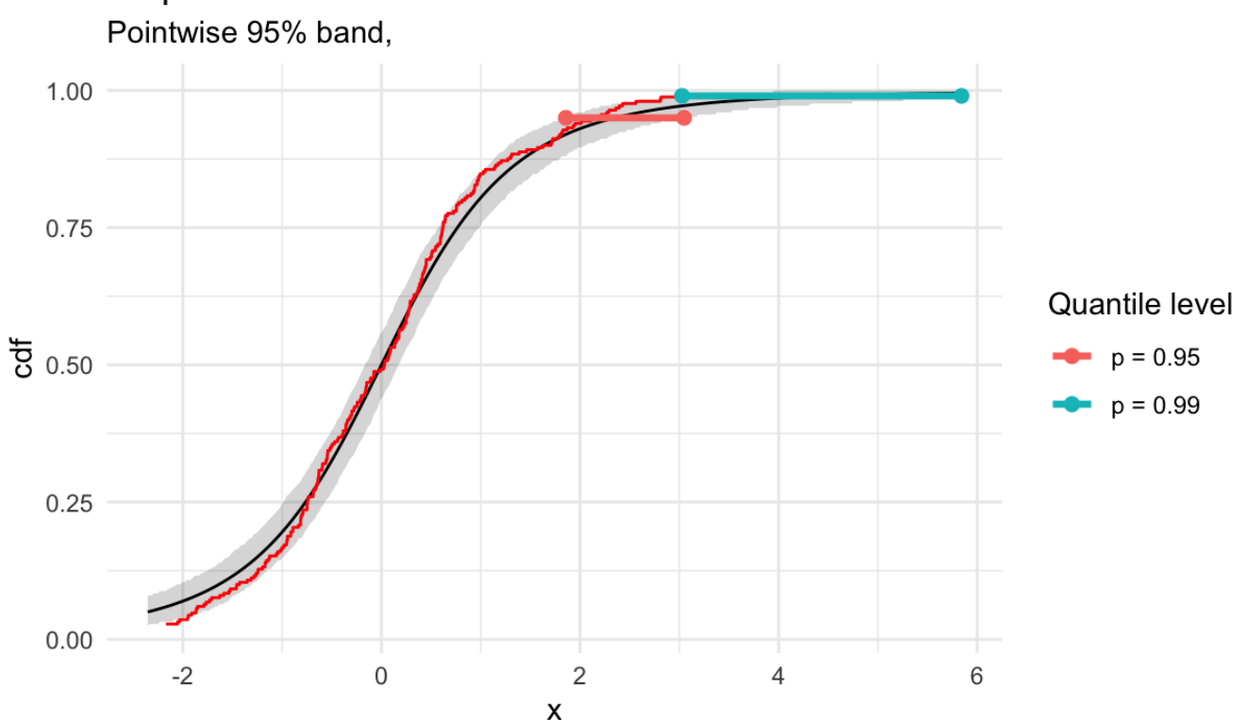
$$P\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \varepsilon\right) \leq 2\exp(-2n\varepsilon^2)$$

- Error between model and unknown truth decays at an exponential rate

Sampling variability of empirical quantiles



Empirical vs true CDF



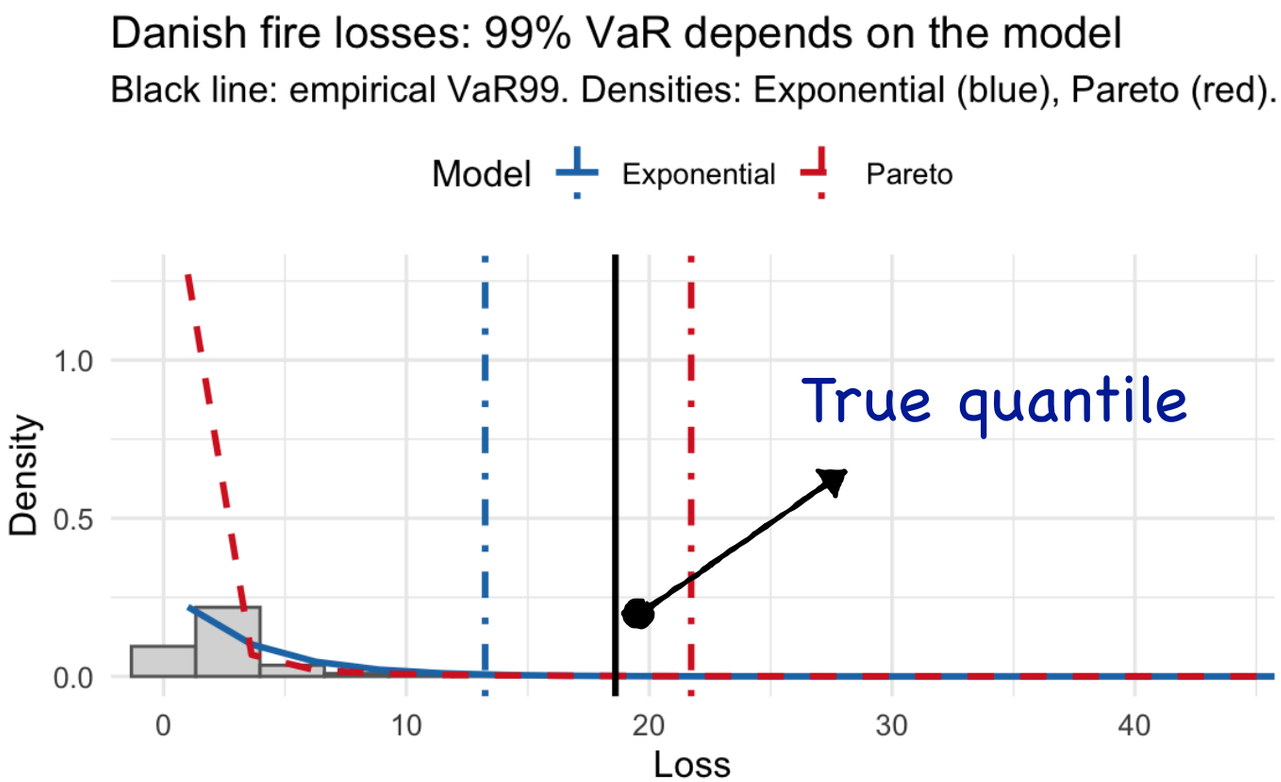
- However, small cdf errors still lead to large errors in tails → use a model

Recourse: Impose (sensible) distributional assumptions

(All models are wrong, some are useful!)

Problem: Suppose you are given losses due to insurance claims and wish to simulate the 99th quantile of the data generating distribution.

Approach: Impose a distributional assumption and simulate



- Exponential → underestimates risk
- Pareto → somewhat more accurate

- Adequacy for features: Model should capture relevant features of data (Not in this course, unfortunately!)
- Simulatability*: It must be easy to generate samples from the model (inverse transforms!)

Agenda for the course

Session 1

Input Modelling 🚂

The bootstrap

Inverse Transforms

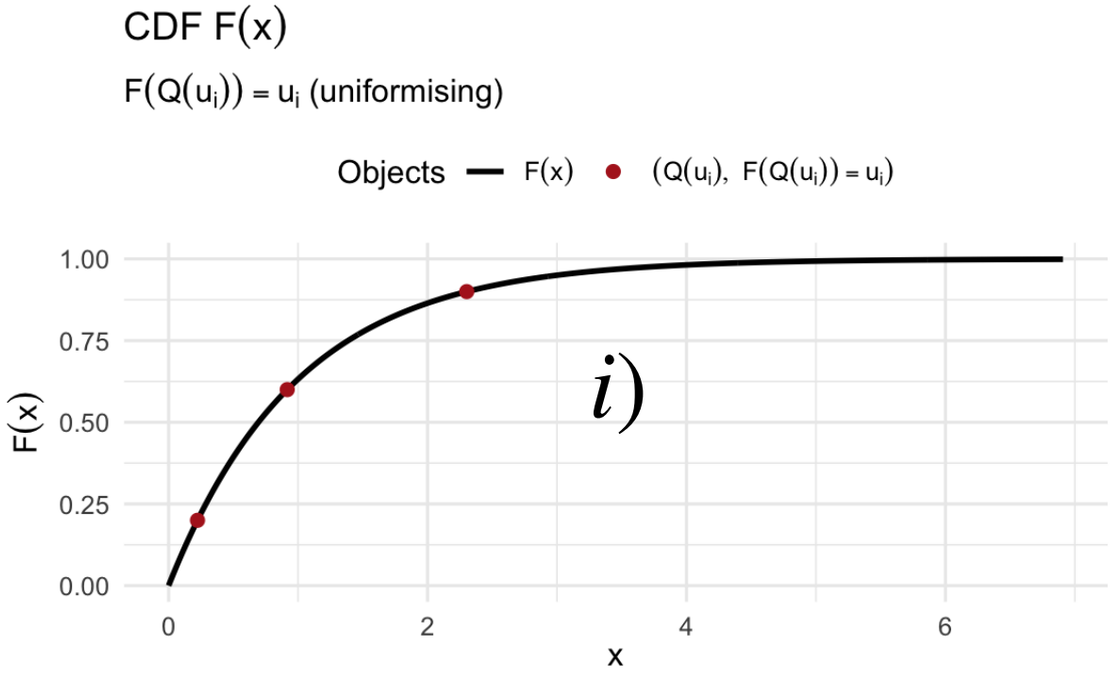
Inverse Transforms: Intuition

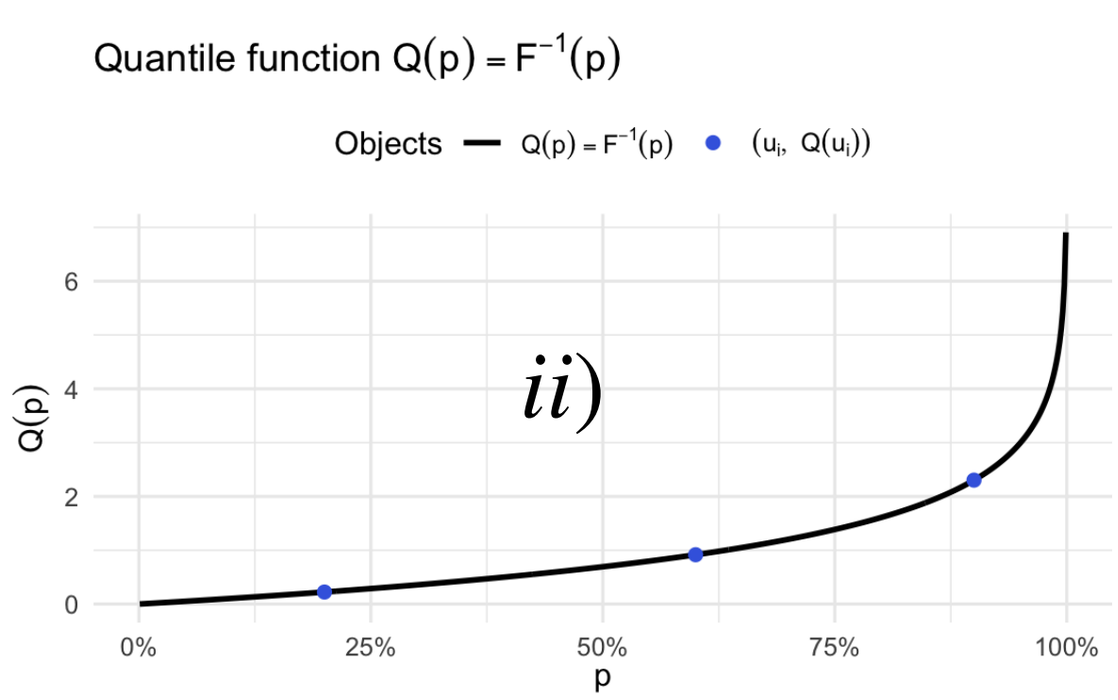
(The quantile function)

Definition: Let $X \sim F$. Then the inverse function of F is called the quantile function of X.

Notation: $Q(p) = F^{\leftarrow}(p)$ where $p \in (0,1]$

Interpretation: $Q(p) \rightarrow$ value below which p fraction of mass lies





Two ways to generate a sample from X

i) Directly sample from X (usually hard)

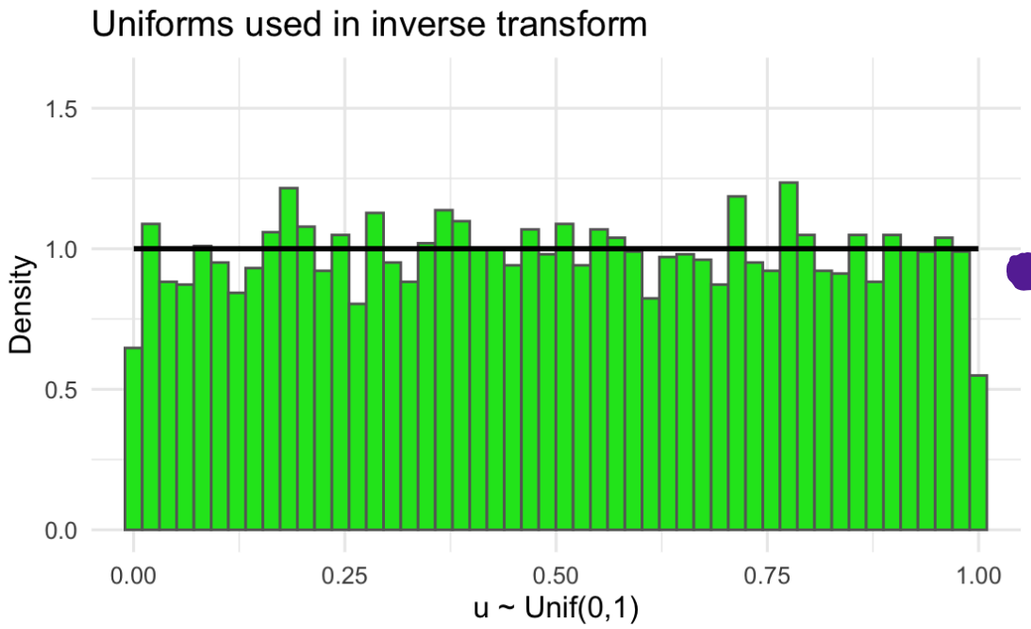
ii) Uniformly draw a percentage and then return its quantile

Inverse Transforms: Method

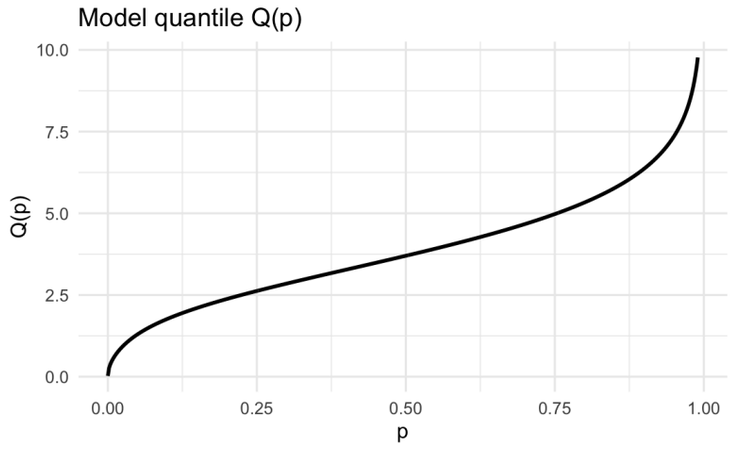
(Mapping distribution —> quantiles)

Task: Let X be a random variable with cdf F. Estimate population parameter θ (e.g. mean, variance, tail probability and so on)

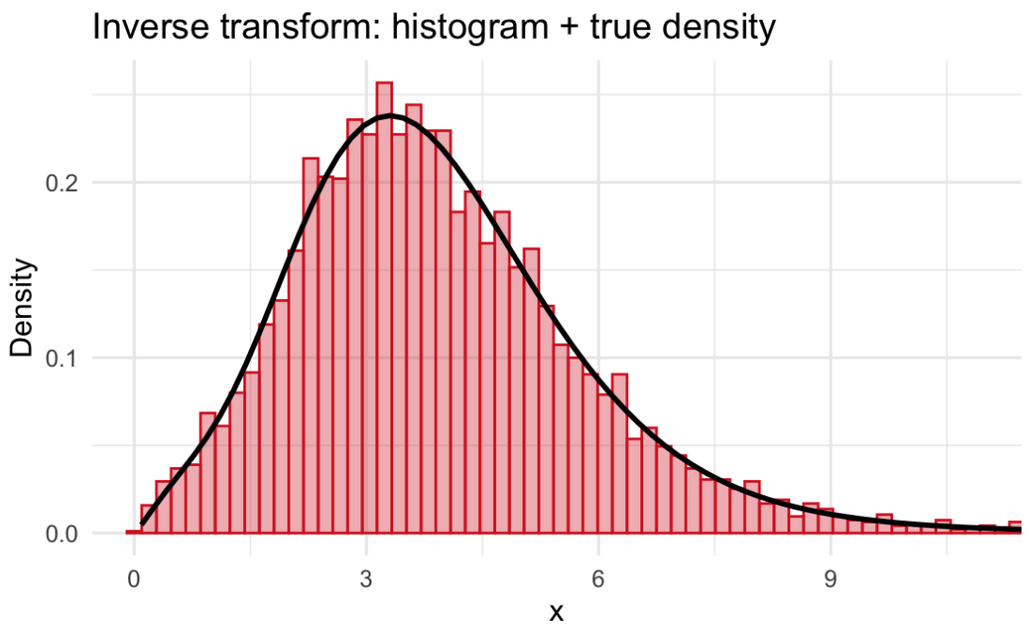
Method: Inverse transforms



$x = Q(u)$



- Step 1: Simulate uniform numbers $u_1, \ldots u_n$
- Step 2: Generate samples $x_i = Q(u_i)$
- Step 3: Return point estimate $\hat{\theta}(x_1, \ldots x_n)$



Inverse Transforms: Why inverse transforms work

(A proof of concept)

Result: Suppose that the quantile function of X, given by $Q(\cdot)$ is monotone increasing function and let U be a uniform random variable on $[0,1]$. Then $Q(U)$ has the same distribution as X.

Proof of inverse transforms: Fix $t \in \mathbb{R}$

- $\mathbb{P}(Q(U) \leq t) = \mathbb{P}(U \leq Q^{-1}(t))$ (since $Q(\cdot)$ is increasing)
- Since U is a standard uniform random variable, $\mathbb{P}(U \leq a) = a$ for $a \in [0,1]$. Thus, the right hand side above equals $Q^{-1}(t)$.
- $F(\cdot)$ is monotone and therefore invertible, so $Q^{-1}(t) = F(t)$. Hence, $\mathbb{P}(Q(U) \leq t) = F(t)$, which suggests $Q(U) \sim X$.

Inverse Transforms: Method

(Mapping distribution –> quantiles)

Result: Suppose that the quantile function of X, given by $Q(\cdot)$ is monotone increasing function and let U be a uniform random variable on $[0,1]$. Then $Q(U)$ has the same distribution as X.

In class example 1: The cdf of a Weibull distribution is given by $1-\exp(-x^k)$. Write down the exact formula that you'd use to generate one sample from a Weibull distribution

Solution:

- For Weibull distribution, $F(x) = 1-\exp(-x^k)$
- This implies that $Q(x) = (\log(1/x))^{1/k}$
- So the algorithm for simulation of one sample becomes
  - Generate $U \sim$ Unif$[0,1]$
  - Output $X = Q(U)$

Inverse Transforms: Method

(Mapping distribution —> quantiles)

Result: Suppose that the quantile function of X, given by $Q(\cdot)$ is monotone increasing function and let U be a uniform random variable on $[0,1]$. Then $Q(U)$ has the same distribution as X.

In class example 2 (Nested Sampling): A bank wishes to simulate the loss given default. Given the economic state x, the borrower defaults with probability $p(x)$. Conditional on default, the rupee loss (in lakhs) follows a Weibull law with $k = 0.4$.

Solution:

- Step 1: Simulate economic state $X \sim P_X$ (if current economic state x is known, then skip this step and set $X = x$)

- Step 2: Simulate $Y = \mathbf{1}(U \leq p(x))$ where $U \sim$ Unif$[0,1]$

- Step 3: Generate $X \sim$ Weibull(k) and output $L = XY$

Agenda for the course

Session 1

Input Modelling 🚂

The bootstrap

Inverse Transforms

Sampling from multivariate distributions

Multivariate Normal Distribution

(The first Multivariate Model)

Task: Recall that bootstrap merely requires access to samples.

- So far: we can only model uniform samples of data
- Broader question —> how to sample from the JOINT distribution of losses?

Basic Model: Multivariate Normal Distribution.

- ξ is said to be multivariate normal if $\xi = \mu + CN$ where N is a d-dimensional vector of independent Gaussians and $C \in \mathbb{R}^{d \times d}$

- It turns out that $E[\xi] = \mu$ and $\text{Cov}(\xi) = \Sigma = CC^\intercal$. Notation: $\xi \sim \text{MVN}(\mu, \Sigma)$

Simulation of a $\text{MVN}(\mu, \Sigma)$ random vector:

- Generate d univariate Gaussian random variables $N = (N_1, \ldots, N_d)$
- Cholesky decomposition: $\Sigma = UDU^\intercal$ where $U \rightarrow$ upper-triangular. Set $C = U\sqrt{D}$
- Return $\xi = \mu + CN$

Multivariate Normal Distribution

(The first Multivariate Model)

Task: Recall that bootstrap merely requires access to samples.

- So far: we can only model uniform samples of data
- Broader question —> how to sample from the JOINT distribution of losses?

Basic Model: Multivariate Normal Distribution.

- ξ is said to be multivariate normal if $\xi = \mu + CN$ where N is a d-dimensional vector of independent Gaussians and $C \in \mathbb{R}^{d \times d}$

- It turns out that $E[\xi] = \mu$ and $\text{Cov}(\xi) = \Sigma = CC^\intercal$. Notation: $\xi \sim \text{MVN}(\mu, \Sigma)$

Proof of Correctness:

- Observe that $N \sim \text{MVN}(0, \mathbf{I})$
- Now $\xi = \mu + CN$ is Gaussian (Gaussianity preserved under linear transforms)
- $E[\xi] = \mu + CE[N] = \mu$ and $\text{Cov}(\xi) = C\,\text{Cov}(N)\,C^\intercal = \Sigma$

Multivariate Normal Distribution

(The first Multivariate Model)

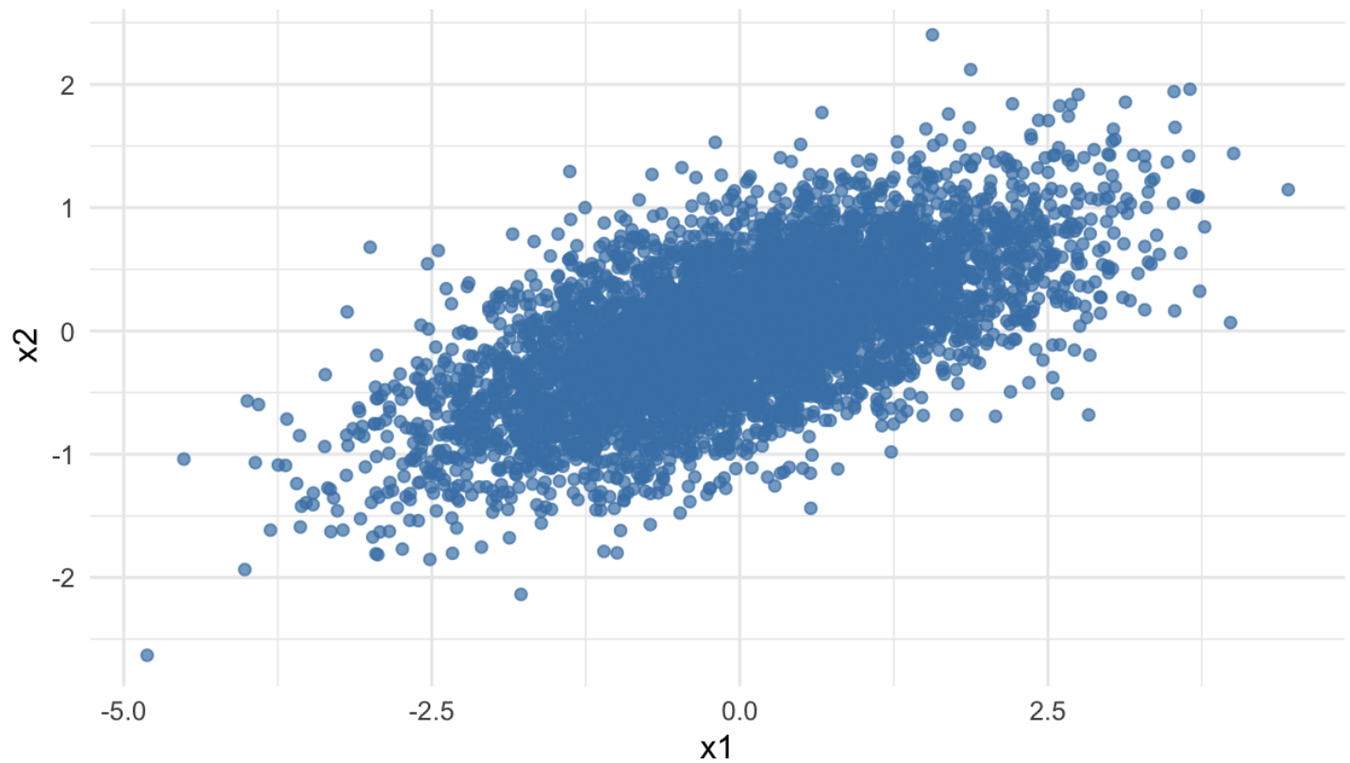
Task: Recall that bootstrap merely requires access to samples.

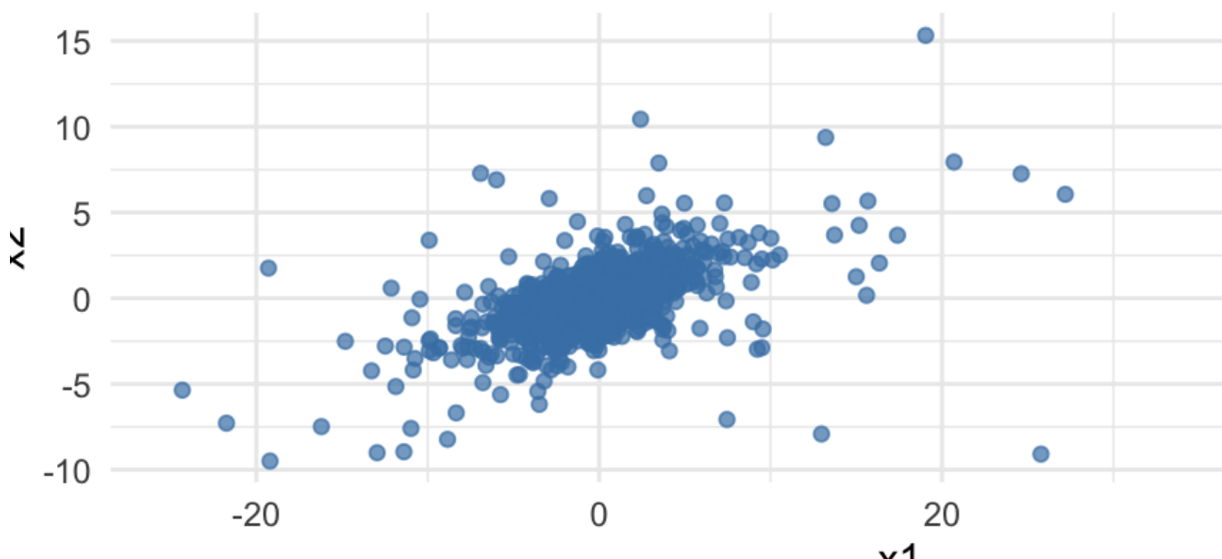
- So far: we can only model uniform samples of data
- Broader question —> how to sample from the JOINT distribution of losses?

- Major Drawback: underestimates joint occurrence of extreme losses (zero tail dependence)

- Use of poor such modelling choices criticised in the for 2008 financial crisis!

- Moral: need more flexible multivariate models!





End of Lecture 1