ECS764P: LECTURE 2

More Probabilities

Dr Fredrik Dahlqvist

Code available at: https://hub.comp-teach.qmul.ac.uk/

THIS WEEK

1. Probability theory and randomness
2. Some important probability measures
3. Centrality: mean, median and mode
4. Dispersion: interquartile range, variance, skewness, kurtosis
5. The pushforward measure

PROBABILITIES AND RANDOMNESS

PROBABILISTIC INTERPRETATIONS

- *There is no randomness in probability theory!*
- *Probability measures just … measure*
- *They assign numbers in* $[0,1]$ *to subsets*
- *These subsets are interpreted as possible outcomes - events*
- *The number is interpreted as a "probability"*

PROBABILISTIC INTERPRETATIONS

- Frequentist *interpretation:*
  - A probability distribution measures the relative frequency of an event as the number of trials/experiments tends to infinity

$$\mathbb{P}(A) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} 1_A(x_i)\,, \qquad (x_i)_{i\in\mathbb{N}} \text{ “sampled from } \mathbb{P}\text{”}$$

- Bayesian *interpretation:*
  - A probability distribution measures the degree to which an agent believes a set of outcomes will come to pass

$$\mathbb{P}(A) = r \;\leftrightarrow \text{I'm } 100r\% \text{ confident } A \text{ will happen}$$

Note: the *indicator function* $1_A(x)$ returns 1 if $x \in A$ and 0 otherwise.

SAMPLING

- If there is no randomness in probability theory, how can we (or Python libraries) **sample** from probability distributions?
- A **Pseudo-Random Number Generator** is a (deterministic!) function capable of outputting numbers which satisfy the frequentist interpretation to an acceptable degree of accuracy.
- A **Hardware Random Number Generator** is a physical device generating random numbers via a physical process (e.g. electromagnetic noise, quantum process) which is known to satisfy the frequentist interpretation to an acceptable degree of accuracy.
- Sampling is usually carried out through a PRNG (sometimes randomly seeded by hardware).

SAMPLING

- A good sampler must satisfy the frequentist interpretation
- Is this enough?
- Consider the probability measure on $\{0,1\}$ defined by $\mathbb{P}(\{0\}) = \mathbb{P}(\{1\}) = \frac{1}{2}$
- Consider the sampler

$$1,0,1,0,1,0,1,0,1,0, \ldots$$

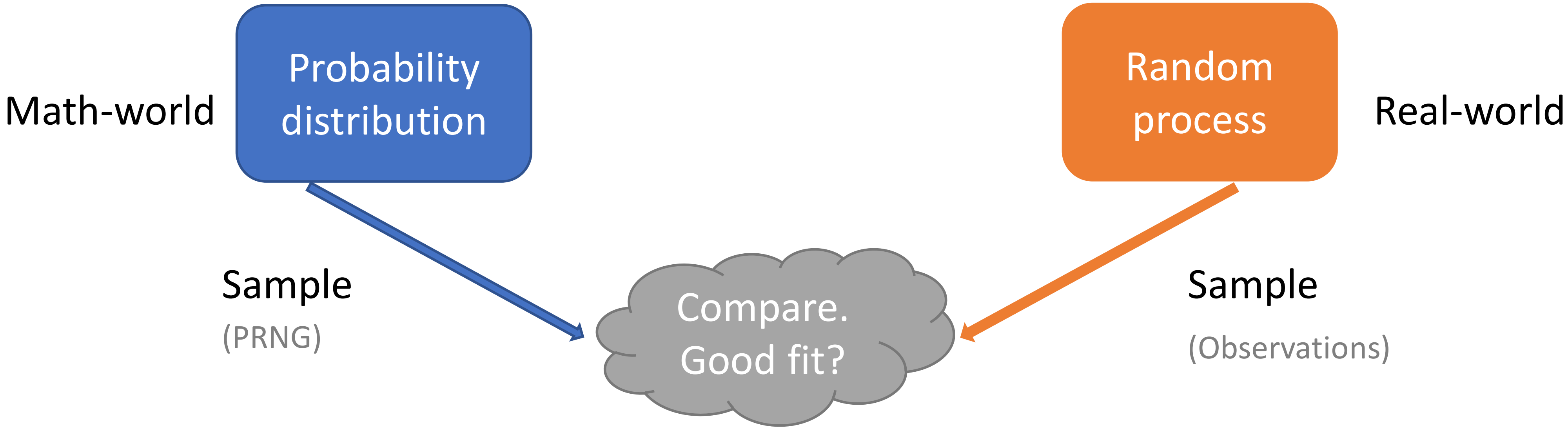
- This satisfies the frequentist interpretation

$$\mathbb{P}(\{0\}) = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} 1_{\{0\}}(x_i) = \frac{1}{2}$$

- And yet … this sampler is manifestly not random!

MODELLING RANDOMNESS

- There might not be any randomness in probability theory, but there exists randomness in our physical world
- We use the frequentist (or Bayesian) interpretation to model sources of randomness



SOME IMPORTANT PROBABILITY MEASURES

THE SUPPORT OF A DISTRIBUTION

Queen Mary
University of London

- Discrete case: given a probability measure $\mathbb{P}$ on a set X, the **support** of $\mathbb{P}$ is the set of elements of X which have non-zero mass

$$supp(\mathbb{P}) = \{x \in X \mid \mathbb{P}(\{x\}) > 0\}$$

- Example: on $\mathbb{N}$ define $\mathbb{P}(\{n\}) = \frac{1+(-1)^n}{2^{n+1}}$
  - Is it a **probability** measure?
  - What is its support?
- Continuous case: harder to formalize. In practice, the set on which the density function is non-zero.

DISCRETE DISTRIBUTIONS WITH FINITE SUPPORT

Queen Mary
University of London

	Dirac Delta	Bernoulli	Binomial	Uniform	Categorical
Notation	δ_x	Bern(p)	Binom(N, p)	Unif(X)	Cat$(p_1, \ldots, p_N)$
Support	$\{x\}$	$\{0,1\}$	$\{0,1, \ldots, N\}$	X finite	$\{0,1, \ldots, N\}$
Parameter(s)	x	$p \in [0,1]$	$N \in \mathbb{N}$, $p \in [0,1]$		$(p_1, \ldots, p_N)$
Density function/PMF	1	$\begin{cases} 1-p & if\ t=0 \\ p & if\ t=1 \end{cases}$	$\binom{N}{k} p^k (1-p)^{N-k}$	$\frac{1}{\lvert X \rvert}$	$f(k) = p_k$

BERNOULLI vs BINOMIAL

- A binomial distribution is a sum of Bernoulli distributions

$$Binom(N, p) = \sum_{i=1}^{N} Bern(p)_i$$

- This makes good intuitive sense, since the sum can be interpreted as the number of successes (i.e. ones)

Note: we will see later what adding probability distributions actually means.

DISCRETE DISTRIBUTIONS WITH INFINITE SUPPORT

	Poisson	Geometric
Notation	Pois(λ)	Geo(N, p)
Support	$\mathbb{N}$	$\mathbb{N}_0$
Parameter(s)	$\lambda \in (0, \infty)$	$N \in \mathbb{N}, p \in (0,1]$
Density function/PMF	$f(n) = \dfrac{\lambda^n e^{-\lambda}}{n!}$	$(1-p)^{N-1}p$

CONTINUOUS DISTRIBUTIONS WITH COMPACT SUPPORT

	Uniform	Beta
Notation	Unif(a, b)	Beta(α, β)
Support	$[a, b]$	$[0,1]$
Parameter(s)	$a < b \in \mathbb{R}$	$a, b \in \ (0, \infty)$
Density function/CDF	$f(x) = \dfrac{1}{b-a}$	$\dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$

Note: Compact means roughly “closed interval” in this context.

DISCRETE DISTRIBUTIONS SUPPORTED BY POSITIVE REALS

	Gamma	χ^2	Lognormal	Exponential	Pareto
Notation	Gamma(α, β)	$\chi^2(k)$	Lognormal(μ, σ^2)	Exponential(λ)	Pareto(x_m, α)
Support	$[0, \infty)$	$[0, \infty)$	$[0, \infty)$	$[0, \infty)$	$[x_m, \infty)$
Parameter(s)	$\alpha, \beta > 0$	$k \in \mathbb{N}_0$	$\mu \in (-\infty, \infty), \sigma > 0$	$\lambda > 0$	$x_m, \alpha > 0$
Density function/CDF	$\frac{\beta^\alpha x^{\alpha-1} e^{-\beta}}{\Gamma(\alpha)}$	$\frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}$	$\frac{e^{\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)}}{x\sigma\sqrt{2\pi}}$	$\lambda e^{-\lambda x}$	$\frac{\alpha x_m^\alpha}{x^{\alpha+!}}$

POISSON vs EXPONENTIAL

- The number of TFL busses arriving at a bus stop in t units of time can be modelled by a Poisson distribution $Pois(\lambda t)$
- Then, the arrival time of the first bus will be distributed according to an exponential distribution $Exponential(\lambda)$.
- Proof: let $\mathbb{P}(A)$ be the probability that the arrival of the first bus is in A

$$\mathbb{P}\big([t,\infty)\big) = Pois(\lambda t)(\{0\}) = e^{-\lambda t}$$

$$\mathbb{P}\big((-\infty,t)\big) = 1 - e^{-\lambda t}$$

- This gives us the CDF. The density is now easily computed

$$f_X(t) = \frac{\partial}{\partial t}\mathbb{P}\big((-\infty,t)\big) = \lambda e^{-\lambda t}$$

DISCRETE DISTRIBUTIONS SUPPORTED BY $(-\infty, \infty)$

	Cauchy	Laplace	Normal	Logistic	Student's t
Notation	Cauchy(x_0, γ)	Laplace(μ, b)	N(μ, σ)	Logistic(μ, s)	Student(n)
Support	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$
Parameter(s)	$x_0 \in \mathbb{R}, \gamma > 0$	$\mu \in \mathbb{R}, b > 0$	$\mu \in \mathbb{R}, \sigma > 0$	$\mu \in \mathbb{R}, s > 0$	$n \in \mathbb{N}_0$
Density function	$\frac{1}{\pi\gamma\left(1+\left(\frac{x-x_0}{\gamma}\right)^2\right)}$	$\frac{e^{\left(-\frac{\lvert x-\mu\rvert}{b}\right)}}{2b}$	$\frac{e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}}{\sigma\sqrt{2\pi}}$	$\frac{e^{\left(-\frac{(x-\mu)}{s}\right)}}{s\left(1+e^{\left(-\frac{(x-\mu)}{s}\right)}\right)^2}$	$\frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\,\Gamma\left(\frac{n-1}{2}\right)\left(1+\frac{x^2}{n-1}\right)^{\frac{n}{2}}}$

MEASURES OF CENTRALITY

Mean, Median and Mode

MEAN

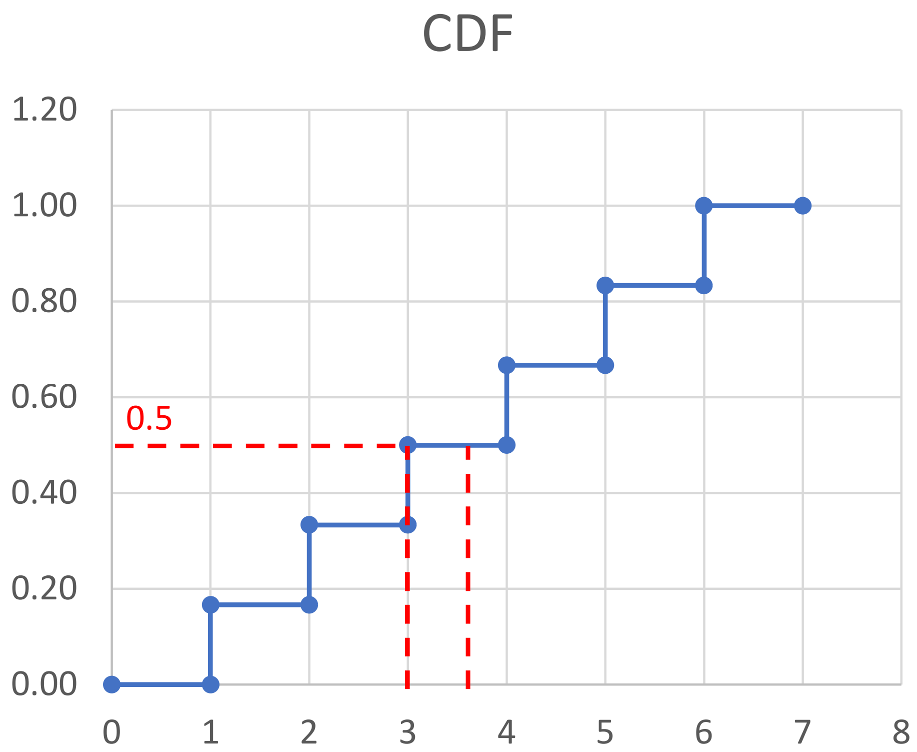
- Recall that probability measures/distributions can be described via their *density functions*
  - Discrete case: reweighting scheme w.r.t. the counting measure
  - Continuous case: reweighting scheme w.r.t. usual integration (dx)
- In what follows we only consider distributions on the real line $\mathbb{R}$
- The **mean** $\boldsymbol{\mu}$ of a probability measure $\mathbb{P}$ with density f is given by
  - Discrete case: $\mu(\mathbb{P}) = \sum_{x:f(x)\neq 0} xf(x)$
  - Continuous case: $\mu(\mathbb{P}) = \int_{-\infty}^{\infty} xf(x)dx$
- Average of all "possible" values, weighted by their "likelihood"

MEAN: EXAMPLES

- Poisson distribution $Pois(\lambda)$ with parameter $\lambda \in (0, \infty)$
  - Density: $f(n) = \frac{\lambda^n e^{-\lambda}}{n!}$ for $n \in \mathbb{N}$, 0 elsewhere
  - Mean: $\mu = \sum_{n=0} nf(n) = \sum_{n=1} \frac{\lambda^n e^{-\lambda}}{(n-1)!} = e^{-\lambda}\lambda \sum_{n=0} \frac{\lambda^n}{n!} = \lambda$
- Uniform distribution $U(a, b)$ on $[a, b]$
  - Density: $f(x) = \frac{1}{b-a}$ for $x \in [a, b]$, 0 elsewhere
  - Mean: $\mu = \int_{-\infty}^{\infty} xf(x)dx = \int_a^b \frac{x}{b-a} dx = \frac{b+a}{2}$

MEDIAN

- The CDF tells us which proportion of the total probability mass lies below a given point t
- Consider the *inverse of the CDF* (inverse CDF): given a probability mass x, it gives every t such the set of all numbers below t weighs x
- Consider for example $x = 0.5$, the inverse CDF gives us t such the set of all numbers below t accounts for half of the total probability
- Not necessarily a function!!!



MEDIAN

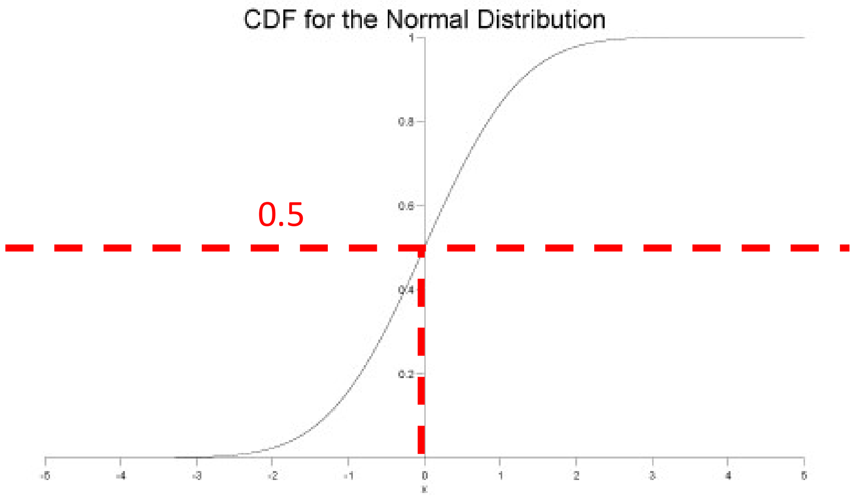
- Formally the median of a distribution with density f is defined as a point m (not necessarily unique!) such that

  - Discrete case:

$$\sum_{x \leq m} f(x) \geq \frac{1}{2} \text{ and } \sum_{x \geq m} f(x) \geq \frac{1}{2}$$

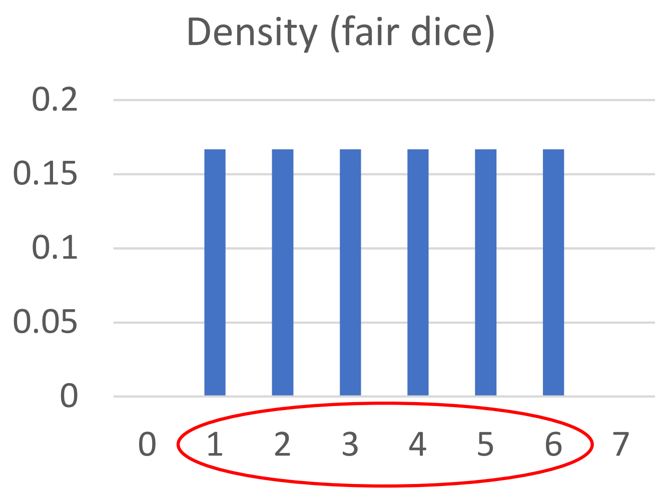
  - Continuous case:

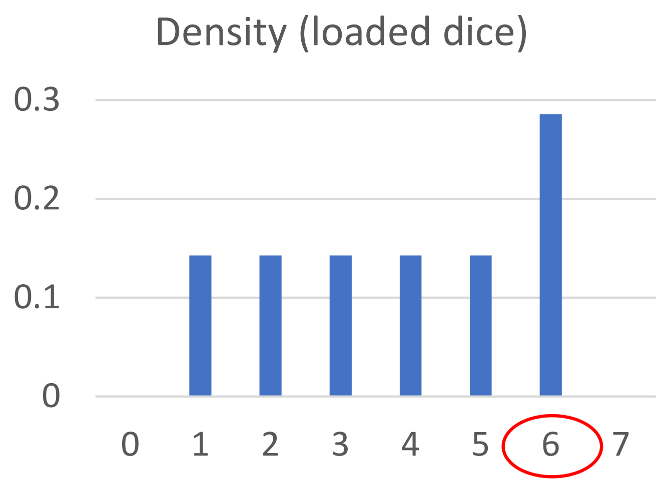
$$\int_{-\infty}^{m} f(x)dx \geq \frac{1}{2} \text{ and } \int_{m}^{\infty} f(x)dx \geq \frac{1}{2}$$



MODE

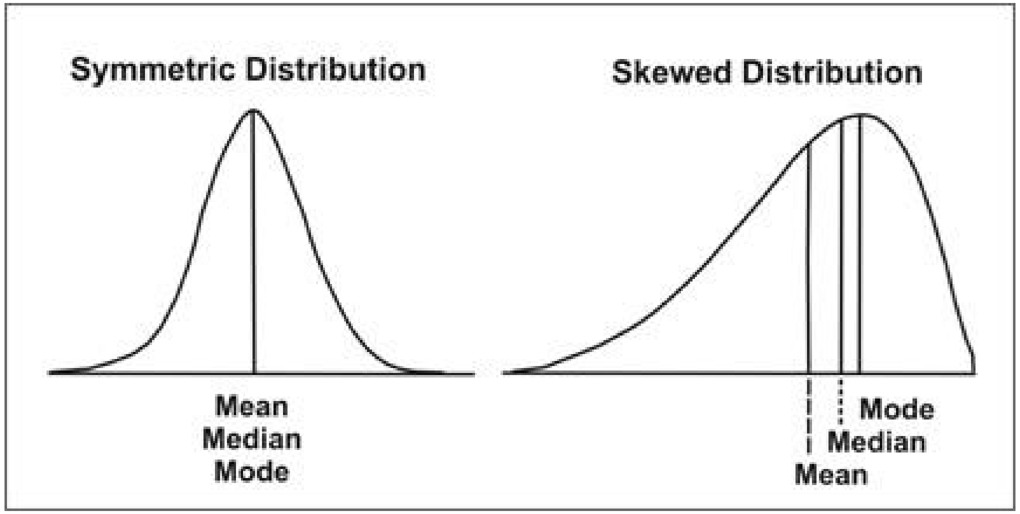
- General idea: value(s) that appear(s) most often when distribution is sampled
- Concretely: point(s) where PMF/PDF is maximal





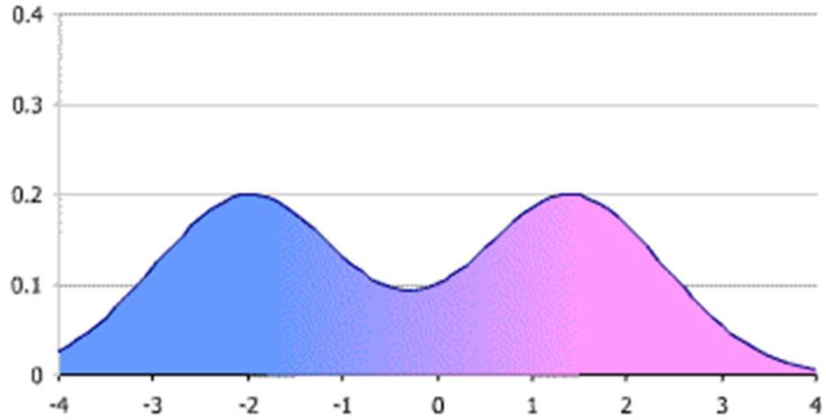
MEAN vs MEDIAN vs MODE

- Difference between Mean, Median and Mode indicate skewed distribution



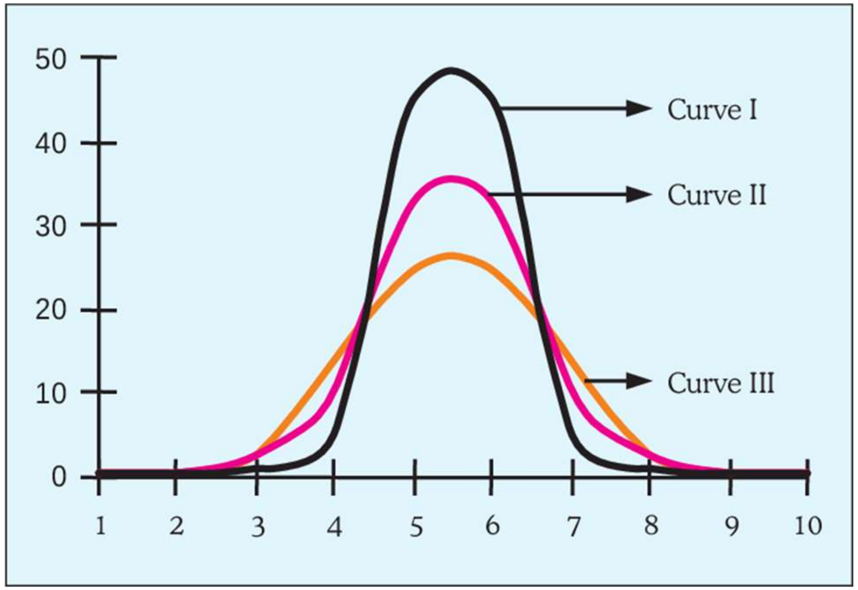
LIMITATIONS OF CENTRAL VALUES

- Typically, same central values for symmetrical distributions.
- Example of exception: distribution with **two maxima**.
  - Mode = any of the maxima.
  - Mean = median = middle value.



LIMITATIONS OF CENTRAL VALUES

- Many distributions can have the same central value, e.g. mean.

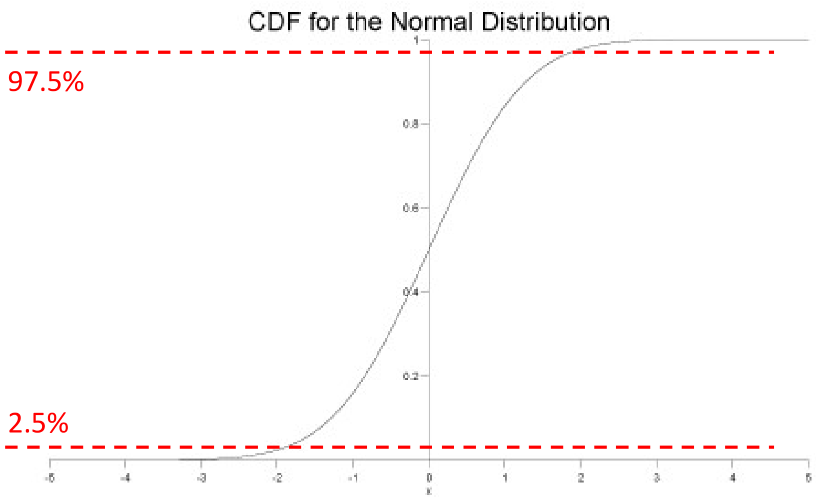


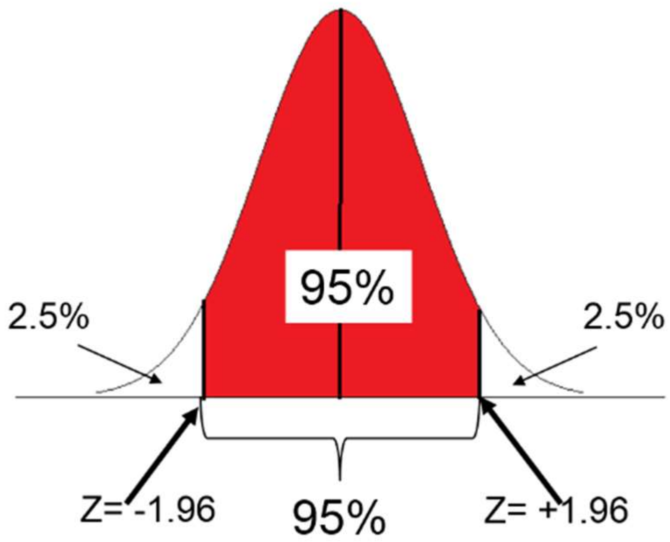
MEASURES OF DISPERTION

Interquartile Range, Variance, Higher-order Moments

PERCENTILE

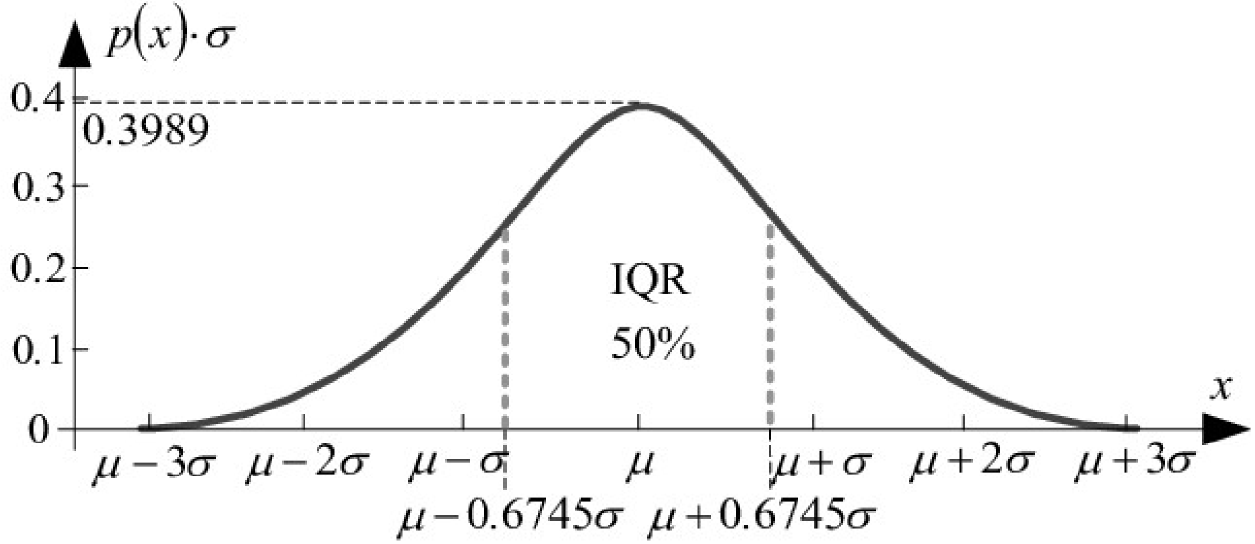
- Remember the inverse of the CDF. Given a probability mass x, it gives t such the set of all numbers below t weighs x





INTERQUARTILE RANGE

- Interquartile range = $75^{th}\ percentile\ - 25^{th} percentile$
- Robust to outliers
- Estimates dispersion



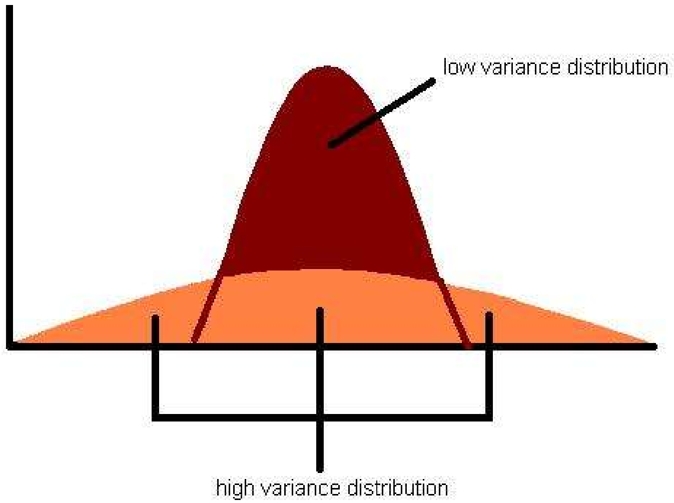
VARIANCE

- The **variance** $\boldsymbol{\sigma^2}$ of a distribution with density f is given by:
  - Discrete case:

$$\sigma^2 = \sum_x (x - \mu)^2 f(x)$$

  - Continuous case:

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$



- Standard deviation σ: square root of variance

VARIANCE vs IQR

- Symmetric, unimodal distributions
  - STD with Mean
- Skewed distribution
  - IQR with Median

SKEWNESS

- The **skewness** $\widetilde{\boldsymbol{\mu_3}}$ of a distribution with density f is given by:
  - Discrete case:

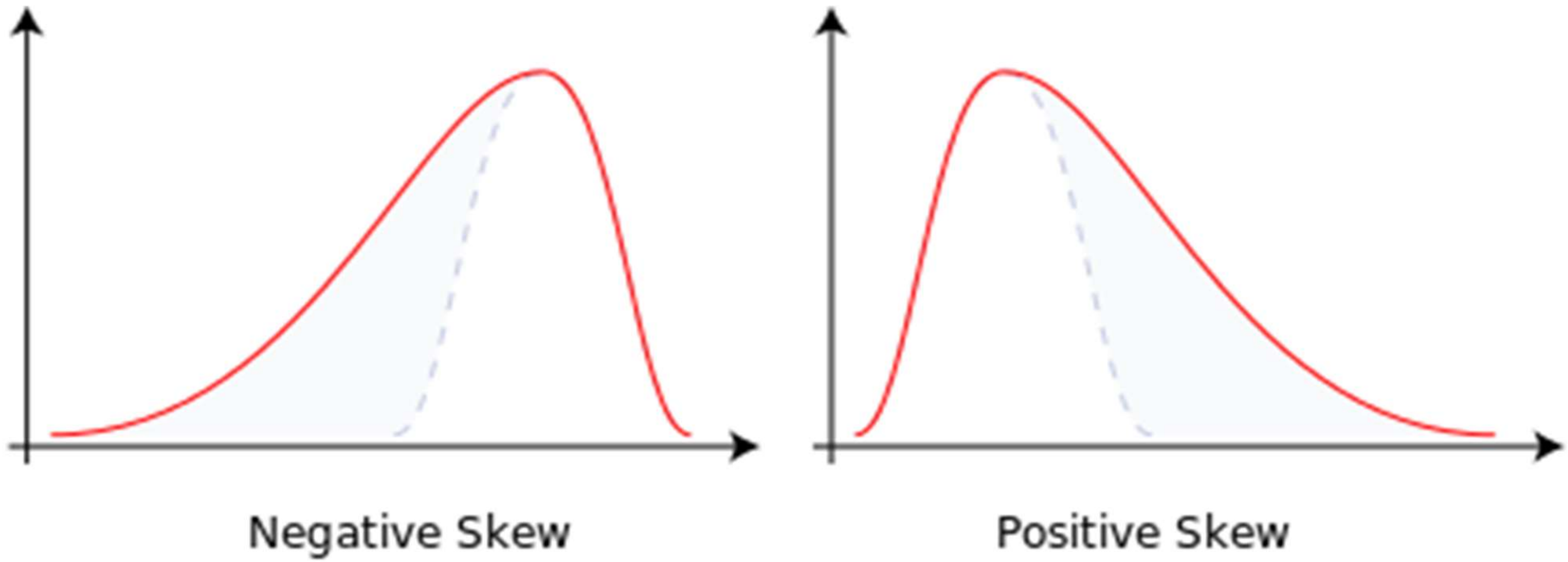
$$\widetilde{\mu_3} = \sum_x \left(\frac{x-\mu}{\sigma}\right)^3 f(x)$$

  - Continuous case:

$$\widetilde{\mu_3} = \int_{-\infty}^{\infty} \left(\frac{x-\mu}{\sigma}\right)^3 f(x)dx$$

- Captures asymmetry

SKEWNESS



KURTOSIS

- The **kurtosis** $\boldsymbol{\widetilde{\mu_4}}$ of a distribution with density f is given by:
  - Discrete case:

$$\widetilde{\mu_4} = \sum_x \left(\frac{x-\mu}{\sigma}\right)^4 f(x)$$

  - Continuous case:

$$\widetilde{\mu_4} = \int_{-\infty}^{\infty} \left(\frac{x-\mu}{\sigma}\right)^4 f(x)dx$$

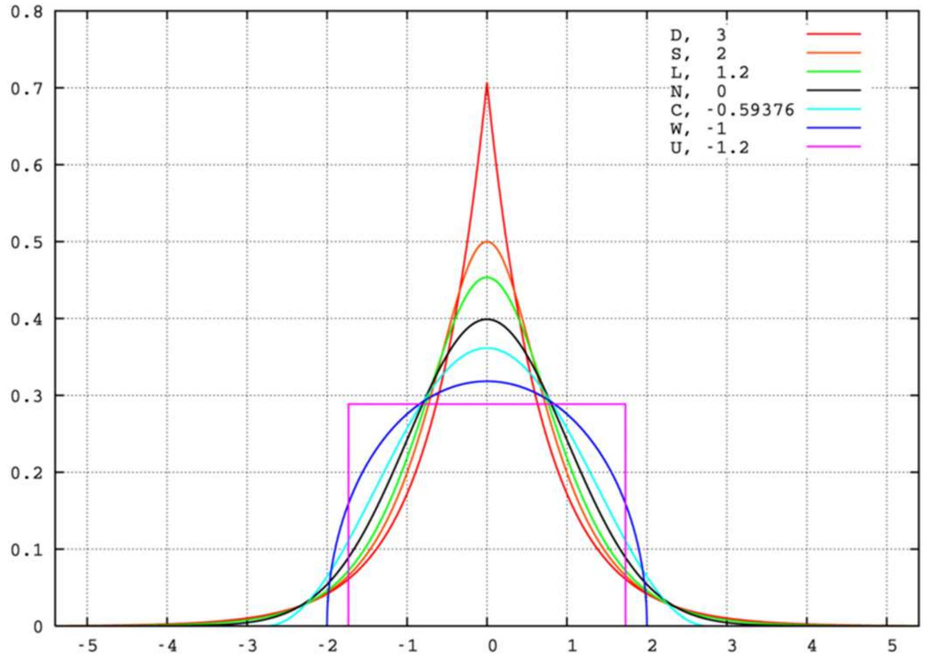
- Captures "peakedness" or "tailedness"

EXCESS KURTOSIS

- The excess kurtosis is defined as:

$$\widetilde{\mu_4} - 3$$

- Three possibilities:
  - Mesokurtic distribution: zero excess kurtosis, i.e. $\widetilde{\mu_4} = 3$
  - Leptokurtic distribution: positive excess kurtosis, i.e. $\widetilde{\mu_4} \geq 3$
  - Platykurtic distribution: negative excess kurtosis, i.e. $\widetilde{\mu_4} \leq 3$



THE PUSHFORWARD MEASURE

APPLYING A FUNCTION TO A MEASURE - THE PUSHFORWARD MEASURE

- Consider the sets $\{1,2,3,4,5,6\}$ and $\{head, tail\}$
- Consider the uniform distribution $\mathbb{P}(\{1\}) = \cdots = \mathbb{P}(\{6\}) = \frac{1}{6}$
- Consider finally, the function

$$f: \{1,2,3,4,5,6\} \to \{head, tail\}, 1,2,3,4 \mapsto head, 5,6 \mapsto tail$$

- Suppose we sample from $\mathbb{P}$ and then apply f, what is the probability of getting $tail$?
- This is given by

$$\mathbb{P}(\{5,6\}) = \mathbb{P}\left(f^{-1}(tail)\right) = \frac{1}{3}$$

Note: for $f: X \to Y, U \subseteq Y$ the inverse image $f^{-1}(U)$ of U is defined as:
$f^{-1}(U) = \{x \in X \mid f(x) \in U\}$

APPLYING A FUNCTION TO A MEASURE - THE PUSHFORWARD MEASURE

- This defines a new probability measure $\mathbb{Q}$ on $\{head, tail\}$ given by

$$\mathbb{Q}(\{tail\}) = \mathbb{P}\left(f^{-1}(tail)\right) = \frac{1}{3}$$

$$\mathbb{Q}(\{head\}) = \mathbb{P}\left(f^{-1}(head)\right) = \frac{2}{3}$$

- $\mathbb{Q}$ is called the pushforward of $\mathbb{P}$ through f and written

$f_*(\mathbb{P})$ or simply $f_*\mathbb{P}$

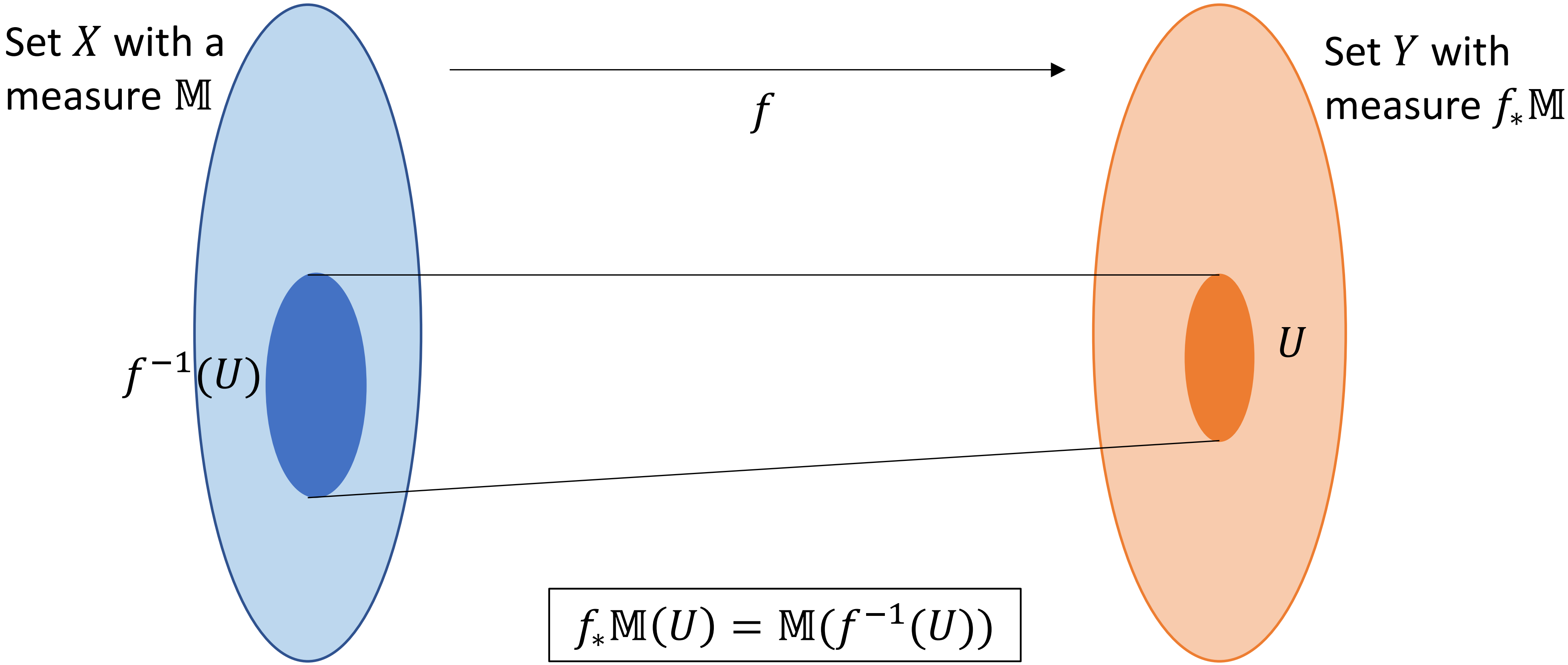
- This is possibly the most important concept of the entire course!

THE PUSHFORWARD MEASURE

- General definition: if $f: X \to Y$ and $\mathbb{M}$ is a measure on X then
- The pushforward of $\mathbb{M}$ through f is the measure on Y defined by

$$f_* \mathbb{M}(U) = \mathbb{M}(f^{-1}(U)) \quad \text{for every } U \subseteq Y$$

THE PUSHFORWARD MEASURE



$$\boxed{f_*\mathbb{M}(U) = \mathbb{M}(f^{-1}(U))}$$

EXAMPLES

- Discrete example:
    - Take $X = Y = \mathbb{N}$ and $\mathbb{M}$ the counting measure
    - Take $f(n) = \left\lfloor \frac{n}{2} \right\rfloor$ (where $\lfloor \quad \rfloor$ is the floor function)
    - $f_*\mathbb{M}(\{1\}) = \mathbb{M}(f^{-1}\{1\}) = \mathbb{M}\left(\left\{n \mid \left\lfloor \frac{n}{2} \right\rfloor = 1\right\}\right) = \mathbb{M}(\{2,3\}) = 2$
    - $f_*\mathbb{M}(\{2\}) = \mathbb{M}(f^{-1}\{2\}) = \mathbb{M}\left(\left\{n \mid \left\lfloor \frac{n}{2} \right\rfloor = 2\right\}\right) = \mathbb{M}(\{4,5\}) = 2$
    - …
    - So $f_*\mathbb{M}$ is just the counting measure multiplied by 2

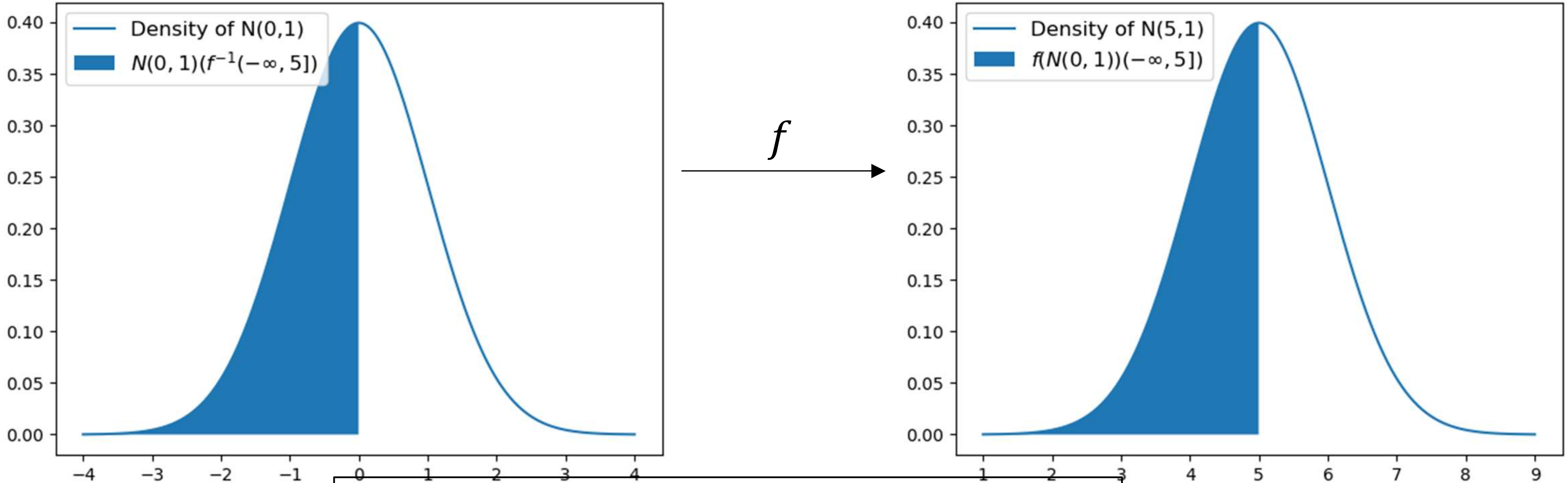
EXAMPLES

- Continuous example:
  - Take $X = Y = [0,1]$ and $\mathbb{P}$ the uniform distribution on $[0,1]$
  - Take $f: [0,1] \to [0,1],\ x \mapsto 1 - x$
  - Let's compute the CDF of $f_* \mathbb{P}$:
  - $f_* \mathbb{P}([0,t]) = \mathbb{P}(f^{-1}[0,1]) = \mathbb{P}(\{x \mid 1 - x \in [0,t]\}) = \mathbb{P}([1-t,1]) = t$
  - Therefore, PDF is 1, in other words $f_* \mathbb{P} = \mathbb{P}$

EXAMPLES

Queen Mary
University of London

Example: $X = Y = \mathbb{R}, f(x) = x + 5$ and $\mathbb{P} = N(0,1)$ standard normal distribution.



$$\boxed{f_*\mathbb{P}((-\infty,5]) = N(0,1)(\{\, x \mid x + 5 \leq 5 \,\}) = \frac{1}{2}}$$

EXAMPLES

Queen Mary University of London

Example: $X = Y = \mathbb{R}, f(x) = x + 5$ and $\mathbb{P} = N(0,1)$ standard normal distribution.

1. CDF computation: $f_*\mathbb{P}\big((-\infty, t]\big) = \mathbb{P}(\{\, x \mid x + 5 \leq t \,\})$

$$= \mathbb{P}((-\infty, t - 5])$$

$$= \int_{-\infty}^{t-5} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx$$

2. PDF computation: differentiate CDF

$$\frac{d}{dt} \int_{-\infty}^{t-5} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-5)^2}{2\sigma^2}}$$

So $f_*\mathbb{P} = N(5,1)$ as expected.