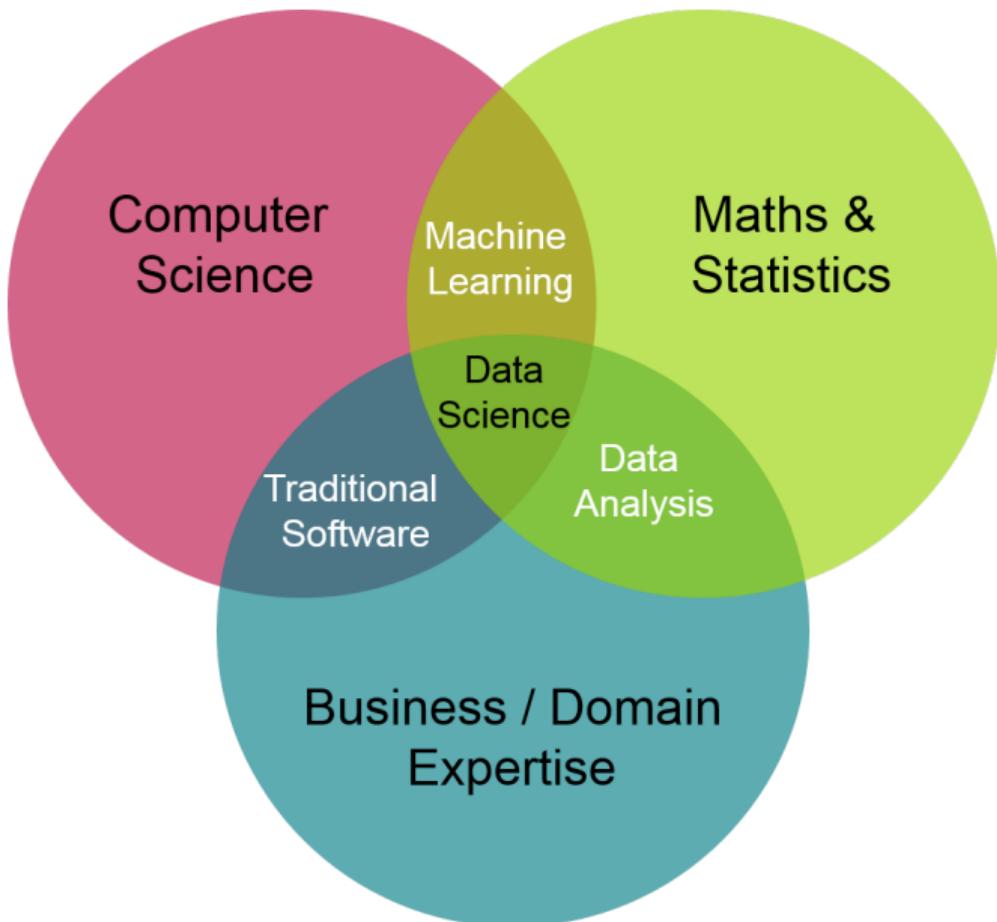


Highdimensional Statistics

Marie Düker

Winter term 2023/24

October 24, 2023



⁰<https://thedataScientist.com/data-science-without-programming/>

Syllabus

- Review: Statistics and Probability
- Motivation: Why high-dimensional statistics? (MW Chapter 1)
- Concentration inequalities (MW Chapter 2 and RV Chapter 2)
- Sparse linear models (MW Chapter 7 and BG Chapter 11)
- Random matrices and covariance estimation (MW Chapter 6)
- Covariance estimation and thresholding (MW Chapter 6 and BL)
- Inverse Covariance estimation (RBLZ)
- Principal component analysis in high dimensions (MW Chapter 8)
- Reproducing kernel Hilbert spaces (MW Chapter 12)
- Review Session

Relevant Literature

- MW: High-Dimensional Statistics: A Non-Asymptotic Viewpoint, by Martin J. Wainwright
- RV: High-Dimensional Probability, by Roman Vershynin
- BG: Statistics for High-Dimensional Data: Methods, Theory and Applications, by Peter Bühlmann and Sara van de Geer
- BL: Covariance regularization by thresholding, by Peter Bickel and Elizaveta Levina
- RBLZ: Sparse permutation invariant covariance estimation, by Adam Rothmann, Peter Bickel, Elizaveta Levina and Ji Zhu

Part 0

Review of some basic concepts in statistics and probability

Outline

- ① Basics
- ② Discrete distributions
- ③ Continuous distributions
- ④ Convergence
- ⑤ Estimators
 - ① General concepts
 - ② The law of large numbers
 - ③ The Central Limit Theorem
- ⑥ Some linear algebra
- ⑦ The multivariate normal distribution

Basics

Experiment A repeatable procedure

① random

② deterministic

- Toss a coin twice
- H head, T tail

Sample space Set of all possible outcomes

$$\Omega = \{HH, TT, HT, TH\}$$

Event Subset of the sample space

$P(\text{at least one tail})$

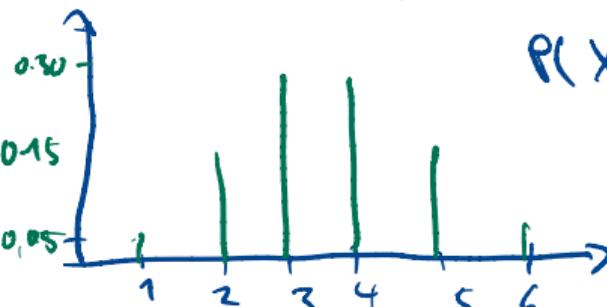
$$= P(\{TT, HT, TH\})$$

$$= \frac{1+TT, HT, TH\}}{4} = \frac{3}{4}$$

Probability function Gives probability for each outcome

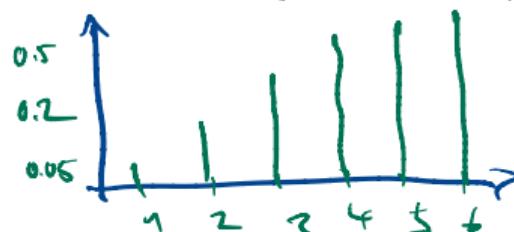
• probability fct.

• 100 students, grades 1-6



$$P(X=3) = 0.3$$

probability distribution $|\Omega|$



Discrete distributions

Discrete random variable Set of possible outcomes is discrete.

Probability mass function The probability mass function (abbreviated pmf) p_X for the random variable X is defined by $p_X(k) = P(X = k)$.

Expected value Given a discrete random variable X which takes values in a set $A = \{x_1, x_2, x_3, \dots\}$, the expected value of X is denoted $E(X)$ or μ_X , and is given by multiplying each possible value of X by its probability.

$$E(X) = \sum_{x \in A} x \cdot P(X = x) = \sum_{x \in A} x \cdot p_X(x).$$

Expected value of $g(X)$

$$E(g(X)) = \sum_{x \in A} g(x) \cdot P(X = x) = \sum_{x \in A} g(x) \cdot p_X(x).$$

Variance of a discrete random variable X is denoted $\text{Var}(X)$, and defined by $\text{Var}(X) = E[(X - E X)^2]$.

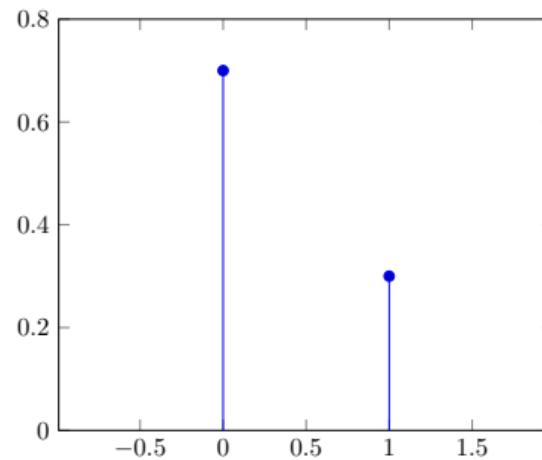
Standard deviation of X is then defined by $\sqrt{\text{Var}(X)}$.

Discrete distributions

Bernoulli The random variable X has a Bernoulli distribution with parameter p , written $X \sim \text{Bernoulli}(p)$, if

$$p_X(k) = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \\ 0 & \text{otherwise.} \end{cases}$$

In other words, X takes the value 1 with probability p , and 0 with probability $1 - p$.



Discrete distributions

If $X \sim \text{Bernoulli}(p)$, then $E(X) = p$ and $\text{Var}(X) = p(1 - p)$.

$$EX = 1 \cdot P(X=1) + 0 \cdot P(X=0) = 1 \cdot p = p$$

$$\text{Var}(X) = E(X - EX)^2 = E(X - p)^2$$

$$= EX^2 - 2EXp + p^2$$

$$\stackrel{\text{def}}{=} pEX$$

$$= EX^2 - p^2$$

*

$$= p - p^2 = p(1 - p)$$

$$* EX^2 = \sum x^2 P(X=x)$$

$$= 1^2 p(X=1) + 0^2 p(X=0)$$

$$= p$$

Discrete distributions

Binomial Distribution The random variable X has a binomial distribution with parameters n and p , written $X \sim \text{Binomial}(n, p)$, if

$$p_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } k = 0, 1, 2, \dots, n \\ 0 & \text{otherwise.} \end{cases}$$

success probability

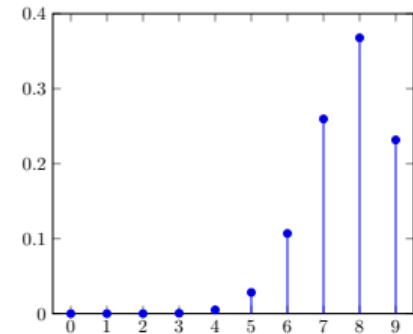
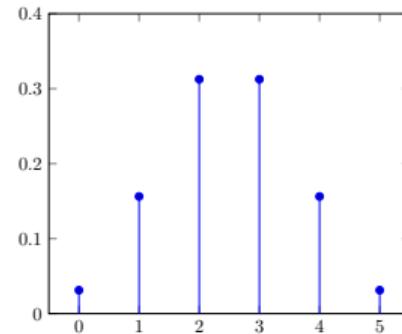
X counts the number of successes after n trials

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad n! = n \cdot (n-1) \cdot (n-2) \cdots 1 \quad \text{Exp: } 3! = 3 \cdot 2 \cdot 1 = 6$$

Exp: S: success, F: failure, success prob. $p=0.9$, 7 attempts

Prob to succeed 4 out of 7 times

$$P(X=4) = \binom{7}{4} 0.9^4 (1-0.9)^3, \quad P(X \leq 4) = \sum_{k=0}^4 P(X=k)$$



Discrete distributions

If $X \sim \text{Binomial}(n, p)$, then $E(X) = np$ and $\text{Var}(X) = np(1 - p)$.

$$\begin{aligned} E(X) &= \sum_{x=0}^n x P(X=x) = \sum_{k=0}^n k P(X=k) \quad \text{Calculate } \text{Var}(X) ! \\ &= \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^n \frac{n(n-1)!}{k!(n-k)!} (1-p)^{n-k} p^{k-n} \\ &= \sum_{j=0}^{n-n} \binom{n-n}{j} p^j (1-p)^{n-n-j} \quad np = np \\ &\quad \underbrace{\qquad\qquad}_{=1} \end{aligned}$$

Discrete distributions

Poisson Distribution The random variable X has a Poisson distribution with parameter λ , written $X \sim \text{Pois}(\lambda)$, if

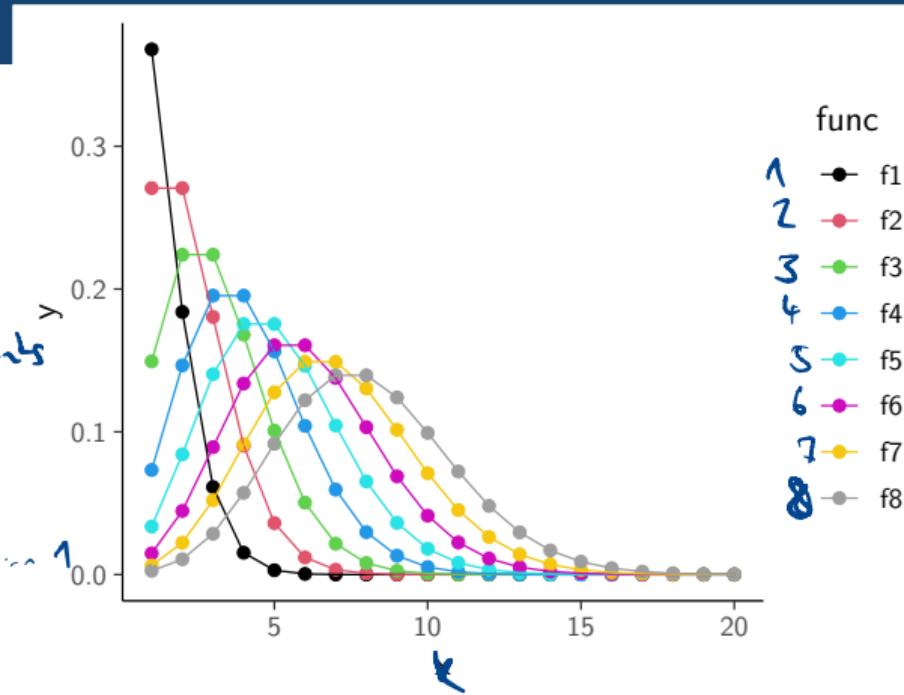
average number of events

$$p_X(k) = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, 2, 3, \dots$$

*number of times
an event occurs*

$$P(X=3) = \frac{2^3 e^{-2}}{3!}$$

$$k! = k \cdot (k-1) \cdots 1$$



Discrete distributions

If $X \sim Pois(\lambda)$, then $E(X) = \lambda$ and $\text{Var}(X) = \lambda$.

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} k \cdot P(X=k) \\ &= \sum_{k=0}^{\infty} k \cdot \frac{e^{-\lambda} \lambda^k}{k!} = \sum_{k=1}^{\infty} \frac{e^{-\lambda} \lambda^k}{(k-1)!} \\ &\quad \underbrace{k(k-1)(k-2)\dots 1}_{k-1=j} \\ &= \sum_{k=1}^{\infty} \frac{e^{-\lambda} \lambda^{k-1} \lambda}{(k-1)!} \stackrel{j=0}{=} \sum_{j=0}^{\infty} \underbrace{\frac{e^{-\lambda} \lambda^j}{j!}}_{=1} \lambda \\ &= \lambda \end{aligned}$$

Continuous distributions

Continuous random variable Takes values in an interval of real numbers.

Probability density function If X is a continuous random variable with cdf F_X , the probability density function of X is denoted f_X and is defined by $f_X(x) = \frac{d}{dx}F_X(x)$. $\int_{-\infty}^{\infty} f_X(x) dx = 1$

Expected value If X is a continuous random variable, the expected value of X is denoted $E(X)$ or μ_X , and defined by the integral

$$E(X) = \int_{-\infty}^{\infty} xf_X(x) dx.$$

Expected value of $g(X)$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f_X(x) dx$$

Variance of a continuous random variable X is denoted $\text{Var}(X)$ and defined by $\text{Var}(X) = E[(X - E X)^2]$.
Standard deviation of X is then defined by $\sqrt{\text{Var}(X)}$.

Some useful properties

For any constants $a, b \in \mathbb{R}$,

1 • $E(aX + b) = aE(X) + b$

2 • $E(X + Y) = E(X) + E(Y)$.

3 • $\text{Var}(X) = E[X^2] - (E[X])^2$

4 • If X and Y are independent, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.

5 • $\text{Var}(aX + b) = a^2 \text{Var}(X)$

1) $E(aX+b) = \int (ax+b)f(x)dx = a \underbrace{\int xf(x)dx}_{=EX} + b \underbrace{\int f(x)dx}_{=1} = aEX + b$

3) $\text{Var}(X) = E(X-EX)^2 = EX^2 - 2E(XEX) + (EX)^2 = EX^2 - 2(EX)^2 + (EX)^2 = EX^2 - (EX)^2$

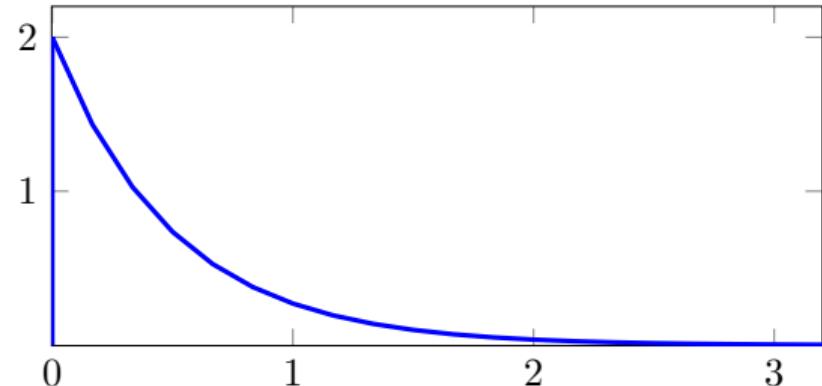
Exercise: Show 2, 4 and 5

Continuous distributions

$$X \sim \text{Exponential}(2)$$

Exponential Distribution The random variable X is exponentially distributed with parameter $\lambda > 0$ (often called the rate parameter), written $X \sim \text{Exponential}(\lambda)$, if its pdf is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$



Continuous distributions

If $X \sim \text{Exponential}(\lambda)$, then $E(X) = \frac{1}{\lambda}$ and $\text{Var}(X) = \frac{1}{\lambda^2}$.

$$\begin{aligned} E(X) &= \int_0^\infty x f(x) dx \\ &= \int_0^\infty x \lambda e^{-\lambda x} dx \end{aligned}$$

$$= \lambda \left[x \left(-\frac{1}{\lambda} \right) e^{-\lambda x} \right]_0^\infty - \int_0^\infty \lambda \left(-\frac{1}{\lambda} \right) e^{-\lambda x} dx \quad *$$

$$= \lambda \left[\left. \frac{1}{\lambda} \int_0^\infty e^{-\lambda x} dx \right] \right] \quad *$$

$$= -\frac{1}{\lambda} e^{-\lambda x} \Big|_0^\infty$$

$$= \frac{1}{\lambda}$$

(*) Integration by parts

$$\begin{aligned} \int_0^\infty g(x) z'(x) dx &= g(x) z(x) \Big|_0^\infty \\ &\quad - \int_0^\infty g'(x) z(x) dx \end{aligned}$$

with $g(x) = x$ and $z(x) = e^{-\lambda x}$

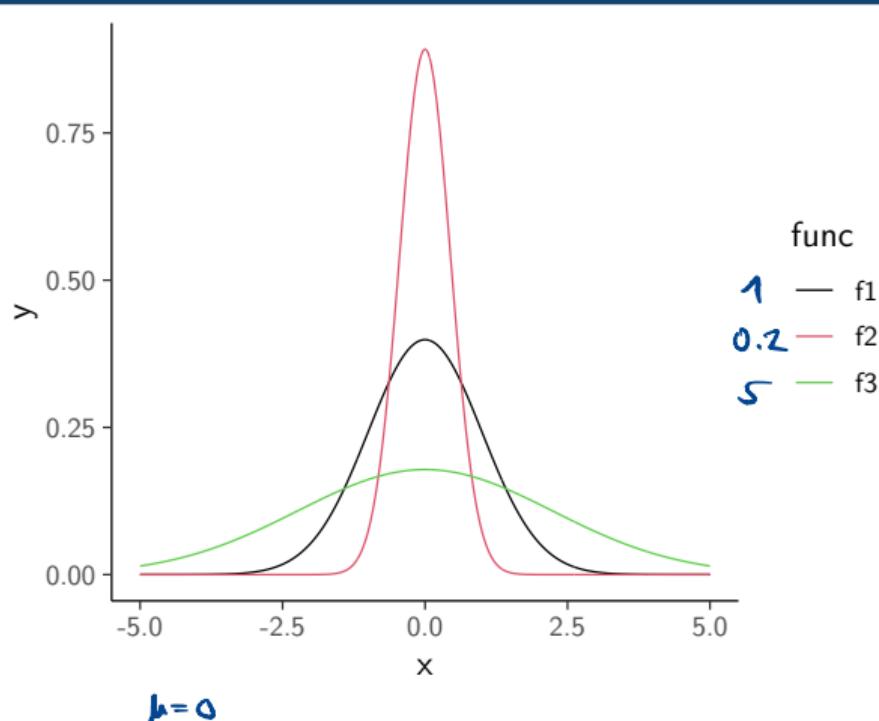
Exercise : $\text{Var}(X) =$



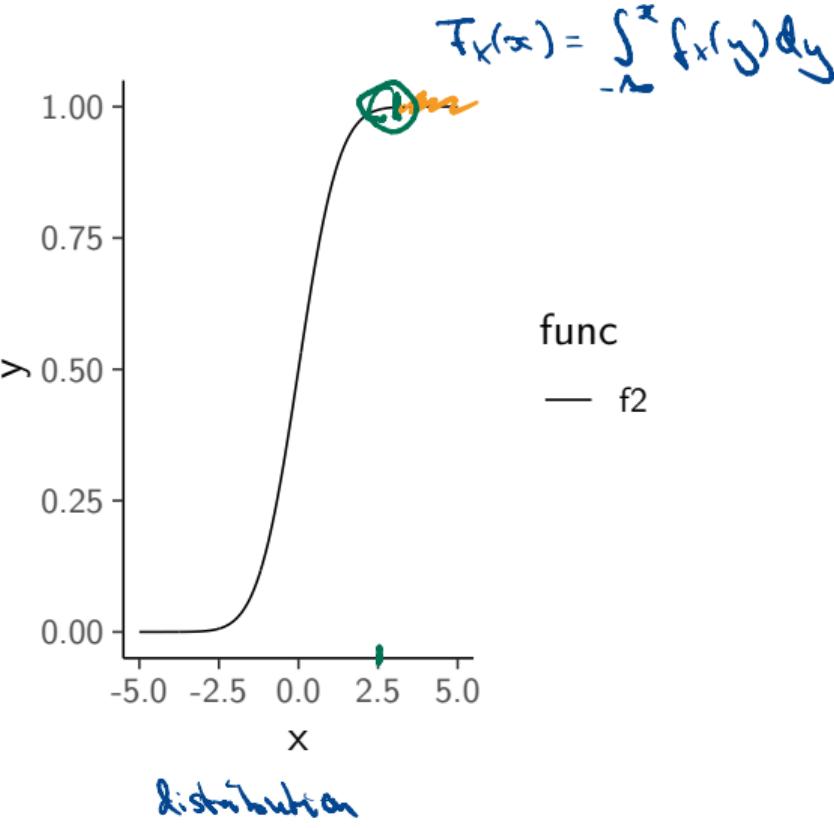
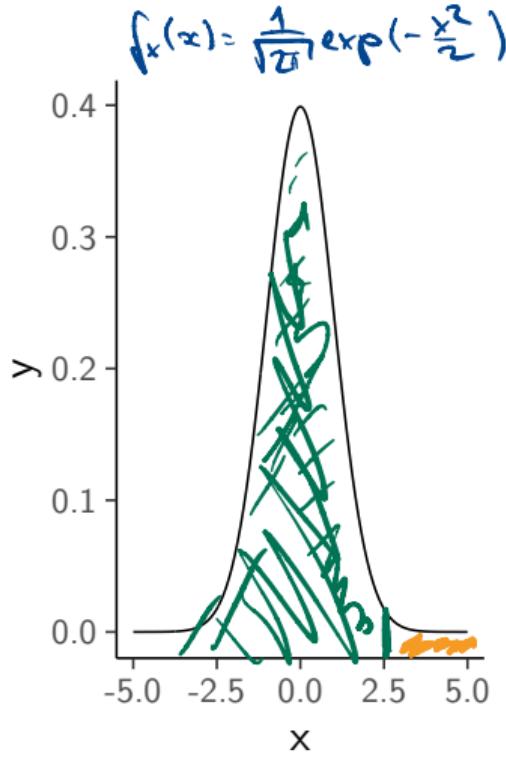
Continuous distributions

Normal/Gaussian distribution The random variable X is normally distributed with parameters μ and $\sigma > 0$, written $X \sim \mathcal{N}(\mu, \sigma^2)$, if

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



Gaussian distribution



Gaussian distribution

Example

- heights of adults is normally distributed.
- heights of adults are normally distributed
 $N(\mu = 175, \sigma^2 = 2.75)$
- Find probability that randomly picked adult is less or equal to 168
 $P(X \leq 168) = \int_{-\infty}^{168} \frac{1}{\sqrt{2\pi\cdot 2.75}} \exp\left(-\frac{1}{2} \frac{(x-175)^2}{2.75}\right) dx$

Some nice property

$$X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2) \text{ and } X_1, X_2 \text{ independent}$$
$$\Rightarrow X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Ex: X_1 : height of male adults,
 X_2 : height of female adults

Estimators

Estimator Rule for estimating a quantity based on observed data.

Bias The bias of an estimator is the difference between its expected value and the actual value of the parameter being estimated.

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

$\text{Bias}(\hat{\theta}) > 0$: $E(\hat{\theta})$ is overestimate of θ , $\text{Bias}(\hat{\theta}) < 0$: $E(\hat{\theta})$ underestimates θ , $E(\hat{\theta}) = \theta$: unbiased estimator

Mean-squared error The mean-squared error of an estimator is the expected value of the squared difference between the estimator and the parameter it estimates.

$$\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2].$$

Example: $\{X_1, X_n\}$ Estimator for the mean $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$:

$$X_i \sim N(\mu, \sigma^2), \text{ Bias } E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n \underbrace{E(X_i)}_{=\mu} = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

$\Rightarrow \hat{\mu}$ is an unbiased estimator

Lemma

$$MSE(\hat{\theta}) = \text{Bias}^2(\hat{\theta}) + \text{Var}(\hat{\theta}), \text{ where } \text{Var}(\hat{\theta}) = E[(\hat{\theta} - E(\hat{\theta}))^2]$$

Proof

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 = E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2] \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) + (E(\hat{\theta}) - \theta)^2] \\ &\stackrel{\text{Linearity } E}{=} \text{Var}(\hat{\theta}) + 2E[(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)] + E[(E(\hat{\theta}) - \theta)^2] \\ &\quad \underbrace{E[\hat{\theta} - E(\hat{\theta})]}_{\text{deterministic}} \underbrace{(E(\hat{\theta}) - \theta)}_{\text{deterministic}} \quad \underbrace{= (E(\hat{\theta}) - \theta)^2}_{= \text{Bias}^2(\hat{\theta})} \\ &\quad = E(\hat{\theta}) - E(\hat{\theta}) = 0 \\ &= \text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta}) \end{aligned}$$

Convergence

Convergence in distribution

A sequence $\{X_n\}$ converges in distribution to a random variable X if

$$\lim_{n \rightarrow \infty} \underbrace{F_n(x)}_{P(X_n \leq x)} = \overbrace{F(x)}^{P(X \leq x)}.$$

Convergence in probability

A sequence $\{X_n\}$ converges in probability to a random variable X if for all $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P(|X_n - X| > \varepsilon) = 0. \Leftrightarrow \lim_{n \rightarrow \infty} P(|X_n - X| \leq \varepsilon) = 1$$

Convergence almost surely

A sequence $\{X_n\}$ converges almost surely to a random variable X if

$$P\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1.$$

Probabilistic O-notation

For random variables X_n and a corresponding set of constants a_n ,

Big O: stochastic boundedness the notation

$$X_n = O_p(a_n)$$

deterministic

means that for any $\varepsilon > 0$, there exists a finite $M > 0$ and a finite $N > 0$ such that

$$P\left(\left|\frac{X_n}{a_n}\right| > M\right) < \varepsilon \quad \text{for all } n > N.$$

$P(|X_n| \leq M |a_n|) \geq \varepsilon$

Small o: convergence in probability the notation

$$X_n = o_p(a_n)$$

means that the set of values $\frac{X_n}{a_n}$ converges to zero in probability. That is,

$$\lim_{n \rightarrow \infty} P\left[\left|\frac{X_n}{a_n}\right| > \varepsilon\right] = 0 \quad \text{for all } \varepsilon > 0.$$

Example

$$\{X_n\}_{n \geq 1} \quad \mu_n = E X_n, \quad \sigma_n^2 = \text{Var}(X_n) \quad < \infty$$

$$X_n - \mu_n = O_p(\sigma_n)$$

$$P(|X_n - \mu_n|/\sigma_n > n^{-\frac{1}{2}})$$

cheatsheet ≠

$$\leq \underbrace{E |X_n - \mu_n|^2 / \sigma_n^2}_{= \sigma_n^2} \frac{1}{n^{-1}}$$

$$= n$$

$$X_n - \mu_n = O_p(s_n) \quad s_n^{-2} \text{Var}(X_n) = \frac{\sigma_n^2}{s_n^2} \rightarrow 0$$

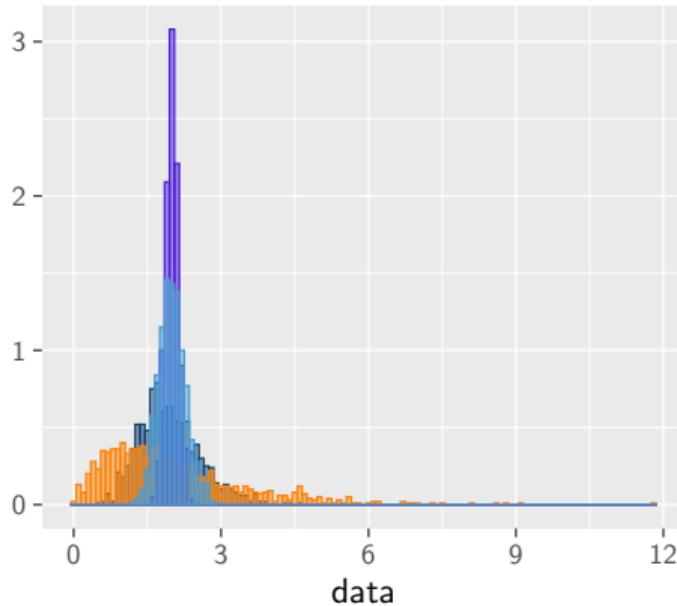
$$P(|X_n - \mu_n| \frac{1}{s_n} > \varepsilon)$$

$$\leq E |X_n - \mu_n|^2 \frac{1}{s_n^2 \varepsilon^2}$$

$$= \frac{\sigma_n^2}{s_n^2} \frac{1}{\varepsilon^2}$$

$$\rightarrow 0$$

CLT



Lemma (CLT)

Let X_1, X_2, \dots be a sequence of i.i.d. random variables with mean μ and variance σ^2 . Consider the sum

$$S_N = X_1 + \dots + X_N \text{ and } Z_N = \frac{S_N - \mathbb{E} S_N}{\sqrt{\text{Var}(S_N)}}.$$

Then, as $N \rightarrow \infty$,

$Z_N \rightarrow \mathcal{N}(0, 1)$ in distribution.



The weak Law of Large Numbers

Lemma (Weak law of large numbers)

Let X_1, X_2, \dots be a sequence of i.i.d. random variables with mean μ . Consider the sum

$$S_N = X_1 + \cdots + X_N.$$

Then, as $N \rightarrow \infty$,

$$\hat{\mathbb{P}}_N S_N \rightarrow \mu \text{ in probability.}$$

$$\forall \varepsilon > 0 \quad \lim_{N \rightarrow \infty} \mathbb{P}(|\frac{1}{N} S_N - \mu| > \varepsilon) = 0$$

The Law of Large Numbers

Lemma (Strong law of large numbers)

Let X_1, X_2, \dots be a sequence of i.i.d. random variables with mean μ . Consider the sum

$$S_N = X_1 + \cdots + X_N.$$

Then, as $N \rightarrow \infty$,

$$S_N \rightarrow \mu \text{ almost surely.}$$

Multivariate random variables

- The multivariate normal distribution is by far the most important multivariate distribution in statistics.
- It's important for all the reasons that the one-dimensional Gaussian distribution is important, but even more so in higher dimensions because many distributions that are useful in one dimension do not easily extend to the multivariate case
- requires some important results from linear algebra

Linear algebra

Inverse The inverse of an $p \times p$ matrix A (if it exists), denoted A^{-1} is the matrix satisfying $AA^{-1} = I_p$.

- Only square matrices have inverses.
- Matrices which are not invertible are called singular

Positive definite A symmetric $p \times p$ matrix A is said to be positive (semi)definite if for all $x \in \mathbb{R}^p$, $x \neq 0$

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \quad x^T Ax > (\geq) 0.$$

Rank The rank of a matrix is the dimension of its largest nonsingular submatrix.

- Note that a nonsingular $p \times p$ matrix has rank p , and is said to be full rank

$$A = \begin{pmatrix} 1 & 2 & | & 3 \\ 4 & 5 & | & 6 \\ \hline 7 & 8 & | & 9 \end{pmatrix}$$
$$\det(A) > 0$$
$$\det\begin{pmatrix} 1 & 2 \\ 4 & 5 \end{pmatrix} = 5 - 8 = -3 < 0$$

Expectation and Variance

X is random vector $X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$ * decorrelation: $A =$

Expectation and variance

$$EX = \begin{pmatrix} EX_1 \\ \vdots \\ EX_p \end{pmatrix}, \quad Cov(X) = E[(X - EX)(X - EX)^T]$$

$$= \left(E[(X_i - EX_i)(X_j - EX_j)] \right)_{i,j=1,\dots,p}$$

Let A be a matrix of constants and X a random vector with mean μ and variance Σ ,

- $E(A^T X) = A^T \mu$ $EATX = A^T EX = A^T \mu$
- * • $\text{Var}(A^T X) = A^T \Sigma A$ $\text{Cov}(A^T X) = E[(A^T X - E(A^T X))(A^T X - E(A^T X))^T]$
 $= A^T E[(X - \mu)(X - \mu)^T] A = (X - \mu)^T A^T A$
- $E(X^T A X) = \mu^T A \mu + \text{tr}(A \Sigma)$,

Linear algebra

$$\text{tr}(A) = \sum_{i=1}^p a_{ii} \quad A = (a_{ij})_{i,j=1,\dots,p}$$

Some useful facts about traces:

- $\text{tr}(AB) = \text{tr}(BA)$
- $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$
- $\text{tr}(cA) = c \text{tr}(A)$
- $\text{tr}(A) = \text{rk}(A)$ if $AA = A$

Some useful facts about the rank:

- A and B with appropriate dimensions, $\text{rk}(AB) \leq \text{rk}(A)$ and $\text{rk}(AB) \leq \text{rk}(B)$.
- $\text{rk}(A^T A) = \text{rk}(AA^T) = \text{rk}(A)$.

Standard normal

$$Z_i \sim N(0, 1) \quad f_{Z_i}(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

Standard normal A real random vector $Z = (Z_1, \dots, Z_p)^T$ is called p -variate standard normal random vector if Z_1, \dots, Z_p are mutually independent and each follows a standard normal distribution.

We write $X \sim \mathcal{N}_p(\mathbf{0}, I)$ and Z has the density:

$$EX = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{p-dim vector}$$

$$f_X(x_1, \dots, x_p) = \frac{1}{\sqrt{(2\pi)^p}} \exp\left(-\frac{1}{2} \underbrace{x^T x}_{\text{sum of squares}}\right).$$

$$(x_1, \dots, x_p) \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} = x_1^2 + \dots + x_p^2$$

Multivariate normal

Multivariate normal A real random vector $X = (X_1, \dots, X_p)^T$ is called a normal random vector if there exists a random p -vector Z , which is a standard normal random vector, a p -vector μ , and a matrix A , such that $X = A^T Z + \mu$.

$\underbrace{p \times p}_{p \times p}$ $\underbrace{p \text{-vector}}$

Suppose $X \sim \mathcal{N}_p(\mu, \Sigma)$ and that Σ is full rank; then X has a density:

$$f_X(x_1, \dots, x_p) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

where Σ denotes the determinant of Σ .

$$E\mathbf{X} = \boldsymbol{\mu}$$

$$\text{Cov}(X) = A^T \underbrace{\text{Cov}(Z)}_{I_p} A = A^T A \stackrel{!}{=} \Sigma$$

Multivariate normal

Properties

- connection between covariance and independence in the multivariate normal distribution
- Suppose $X \sim \mathcal{N}_p(\mu, \Sigma)$. If $\Sigma_{ij} = 0$, i.e. the off-diagonal corresponding to X_i and X_j is zero, then X_i and X_j are independent.
- linear combinations are also normally distributed
- Let b be a $p \times 1$ vector of constants, B a $p \times p$ matrix of constants, and $X \sim \mathcal{N}_p(\mu, \Sigma)$. Then

$$b + BX \sim \mathcal{N}_p(B\mu + b, B\Sigma B')$$

$$- E(b + BX) = E(b) + E(BX) = b + E(BX) = b + BEX = b + B\mu$$

distributiv
gesetz

$$- E[(b + BX - E(b + BX))(b + BX - E(b + BX))^T]$$

$$= E[(b + BX - (b + B\mu))(b + BX - (b + B\mu))^T]$$

$$= E[(BX - B\mu)(BX - B\mu)^T] = B E[(X - \mu)(X - \mu)^T] B^T = B \Sigma B^T$$

Summary

- What is the difference between discrete and continuous distributions?
- How does one calculate the expected value and the variance of a random variable?
- What are some basic properties of the expected value and the variance?
- How is convergence in probability defined?
- How is convergence in distribution defined?
- What do the CLT and LLN tell us?
- What is the probabilistic big-O notation?
- How to calculate mean and covariance of multivariate normal distribution?
- How is the rank of a matrix defined?