

# Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#211 Statistical Functions (cdf and density)

été 2020

## Statistical Functions

gradum inter se eque distantium. ¶ **L**atitudinem unius inter diffinitionem quendam sinit ut a non gradu a certum gradum: quia line per a certo gradum a certum gradum. Pro eni pot bar latitudo: incipit a non gradu a certum ad a gradum quod est uniusmodi diffinitionis: quod in puncto intendit et in fine remittit huiusmodi diffinitionis semper per intentionem. ¶ **L**atitudinem diffinitionis diffinitionis quod per se tota est diffinitionis diffinitionis: quia huiusmodi latitudo per se tota diffinitionis diffinitionis est illa: cuius nulla pars est uniusmodi aut uniusmodi diffinitionis aut e converso. ¶ **L**atitudo non per se tota diffinitionis diffinitionis: est cuius aliqua pars est diffinitionis siue uniusmodi diffinitionis. ¶ **L**atitudinem diffinitionis diffinitionis scilicet est tota quod aut uniusmodi diffinitionis diffinitionis et quod diffinitionis inter diffinitionis diffinitiones. Pro quod non dicitur quod sicut ymaginamur latitudinem in nulla in puncto et variata quod vocamus uniusmodi. Quandoque in suis partibus variam quod vocamus diffinitionem tantum. Quandoque huiusmodi variatur: vocatur uniusmodi diffinitionis. Si vero diffinitionis variatur vocatur diffinitionis diffinitionis: ita ymaginamur quodam variatorem latitudinem uniusmodi quodam diffinitionem. Et rursum variatorem diffinitionem quodam diffinitionis diffinitionem et quodam diffinitionis diffinitionis diffinitionem. ¶ **I**nde sicut uniusmodi latitudinis varia: reddidit uniusmodi inter diffinitionis diffinitionem. Ita

diffinis vniſomiter var'atio reddit vniſo-  
 miter diſſomiter diſſormis. ¶ **Quia** vni-  
 ſorm e diſormis e illo q' in e cellis gradus  
 e q' diſanus fuit eide p'portio a la t'ia p-  
 portio e q'latia. Tia ſi vni e cellis gradus  
 inter e e q' diſanus fuerit p'portio e q'la-  
 tis t'ie e e la t'ia vniſormis diſormis ut p' e  
 diſſimilioribus membro ſecid' diſſormis  
 Rarſus ſi nulla proportio ſeruat tunc nulla  
 poſſet attendi vniſormitas in latitudine tali  
 ſic non eſſe vniſomiter diſo. m e diſſormis.  
 ¶ **Quia** vniſormiter diſormiter diſſormis  
 e illo q' inter e cellis gradus eque diſſanum  
 non erit eandem proportionem ſicu. n ſe  
 cunda parte patebit. Notandum tamen eſt  
 q' ſicut in ſupradictis diſſimilit' ubi loquor  
 de eceſſu graduum inter e eque diſſanum  
 debz accipi diſſantia ſecdm parces latitudinis  
 eſſentia t' no inſineſſentia ut loquunt' de cie  
 ſimilit' de diſſana q' dui ſimili ſi aut graduale

.....

... 2002-03-01

Nicole Oresme, *Tractatus de latitudinibus formarum*, 1486

# Cumulative Distribution Function

Given a random variable  $X$ ,  $F(y) = \mathbb{P}[X \leq x]$

$F$  is an increasing function, taking values in  $[0, 1]$ .

Consider a sample  $\mathbf{x} = \{x_1, y_2, \dots, x_n\}$ , a natural estimator is

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(x_i \leq x)$$

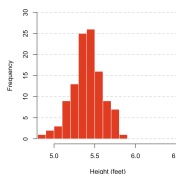
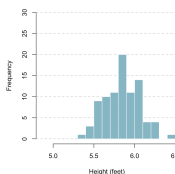
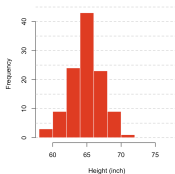
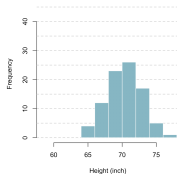
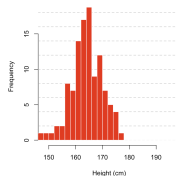
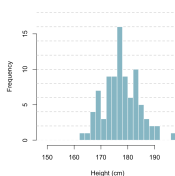
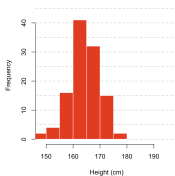
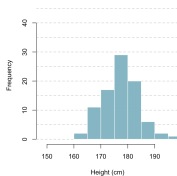
```
1 > import numpy
2 > x = numpy.sort(x)
3 > n = x.size
4 > y = numpy.arange(1, n+1) / n
5 > import matplotlib.pyplot as plt
6 > plt.plot(x, np.linspace(0, 1, n, endpoint=False))
```

```
1 > x = sort(x)
2 > n = length(x)
3 > y = (1:n)/n
4 > plot(ecdf(x))
```

# Density & Histogram

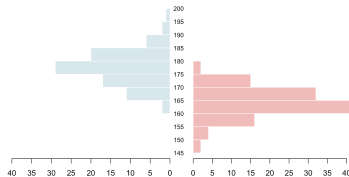
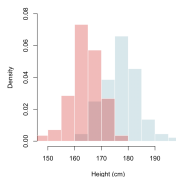
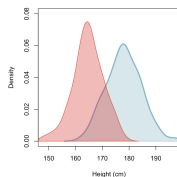
Given a random variable  $X$ ,  $f$  is such that  $F(x) = \int_{-\infty}^x f(t)dt$   
or conversely,  $f(x) = F'(x)$ .

Thus,  $\mathbb{P}(X \in [a, b]) = \int_a^b f(t)dt$



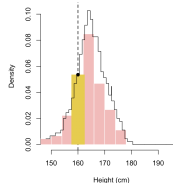
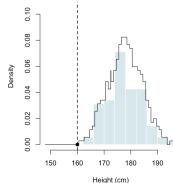
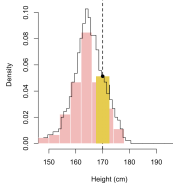
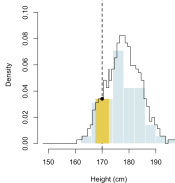
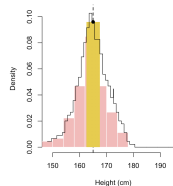
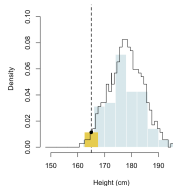
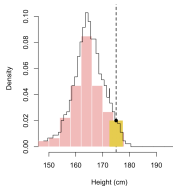
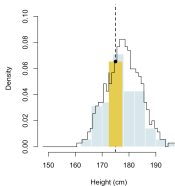
# Histogram & Density

Can be used to compare distributions



# Moving Histogram

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n \mathbf{1}(|x_i - x| \leq h/2)$$



# Moving Histogram

$\hat{F}$  cannot be differentiated, but we can consider

$$f_h(x) = \frac{1}{h} \underbrace{F(x + h/2) - F(x - h/2)}_{\mathbb{P}(X \in [x \pm h/2])}$$

i.e.

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n \mathbf{1}(x_i \in [x - h/2, x + h/2])$$

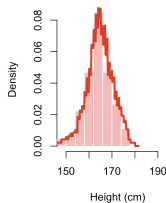
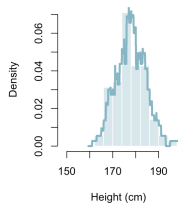
$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n \mathbf{1}(|x_i - x| \leq h/2)$$

One can prove that  $\mathbb{E}(\hat{f}_h(x)) = f_h(x) \sim f(x) + \frac{h^2}{24} f''(x)$

i.e.  $\text{bias}(\hat{f}_h(x)) \sim \frac{h^2}{24} f''(x)$ , while  $\text{Var}(\hat{f}_h(x)) \sim \frac{1}{nh} \cdot f_h(x)$

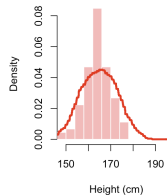
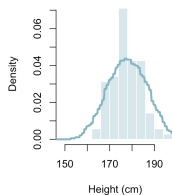
# Moving Histogram

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n \mathbf{1}(|x_i - x| \leq h/2)$$



small  $h$

bias  $\text{bias}(\hat{f}_h(x))$  small  
variance  $\text{Var}(\hat{f}_h(x))$  large



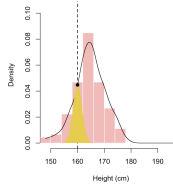
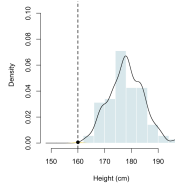
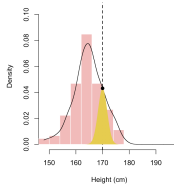
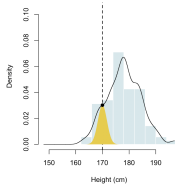
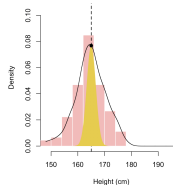
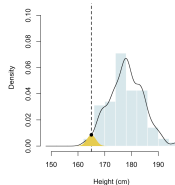
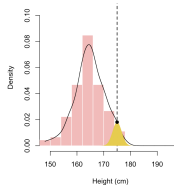
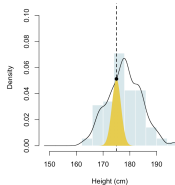
large  $h$

bias  $\text{bias}(\hat{f}_h(x))$  large  
variance  $\text{Var}(\hat{f}_h(x))$  small



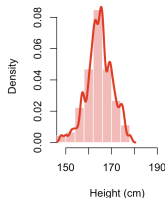
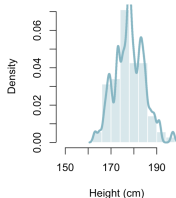
# Kernel Density

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)$$



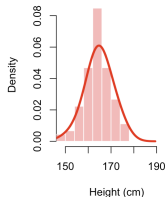
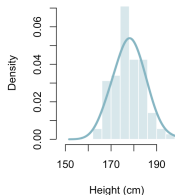
# Kernel Density

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)$$



small  $h$

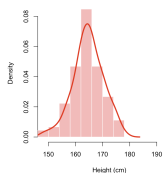
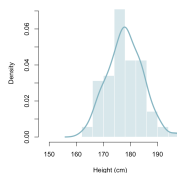
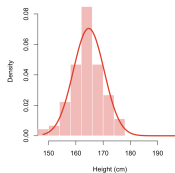
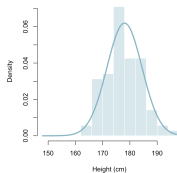
bias  $\text{bias}(\hat{f}_h(x))$  small  
variance  $\text{Var}(\hat{f}_h(x))$  large



large  $h$

bias  $\text{bias}(\hat{f}_h(x))$  large  
variance  $\text{Var}(\hat{f}_h(x))$  small

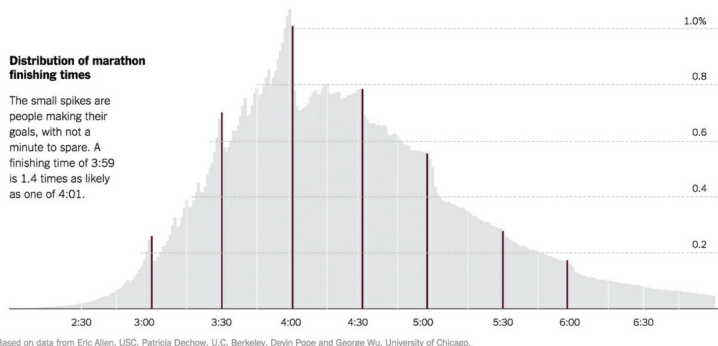
# Histogram & Density



```
1 > import matplotlib.pyplot as plt
2 > hist = plt.hist(x, bins=30, normed=True)
3 > from sklearn.neighbors import KernelDensity
4 > k = KernelDensity(bandwidth=1.0, kernel='gaussian')
5 > k.fit(x[:, None])
```

```
1 > hist(x, probability=TRUE)
2 > plot(density(x))
3 > plot(density(x), kernel="gaussian", bw=1)
```

# Histogram & Density



## Reference-Dependent Preferences: Evidence from Marathon Runners