probability vs. statistics

~~support~~ →

→ reduce dimensionality

$$\underbrace{[x_1, x_2, \dots, x_{10}]}_{\text{mean}([\,])}$$

1 #

PROBABILITY

random variables: X, Y

likelihood of events in a $\underbrace{\text{sample space}}$

you are going to define

situation: coin flip

$S = \{H, T\}$

$= \{H, T, BOOE\}$

H	T

$P(H) = 0.5$
$P(T) = 0.5$

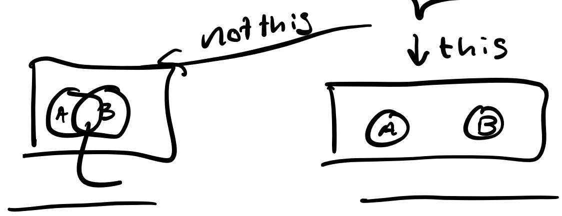
H	T

$P(H) = 0.49995$
$P(T) = 0.49994$
$P(BOOE) = 0.0001$

Axioms of Probability

1) nonnegativity
2) normalization $\sum_x P(x) = 1$
3) additivity

if independent : $A \cap B = \emptyset$



$$P(A \cap B) = P(A) + P(B)$$

suggestions for SS

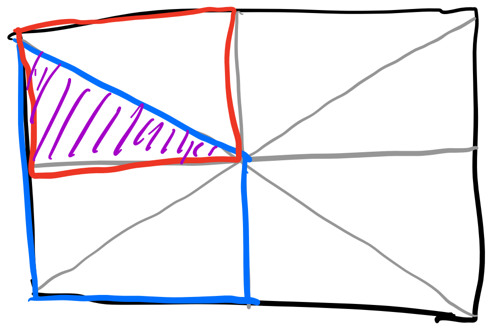
1. mutually exclusive
2. collectively exhaustive

} not always possible

conditional probability

$P(A \mid B)$: "probability of A happening given B happened"

$P(A) = 2/8$



$P(A \cap B) = 1/8$

$P(A \cup B) = 1/2$

$P(B) = 3/8$

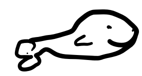
$$P(A \mid B) = \frac{1/8}{3/8} = 1/3$$

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

joint probability

$$P(A \cap B) = P(A)\, P(B \mid A)$$

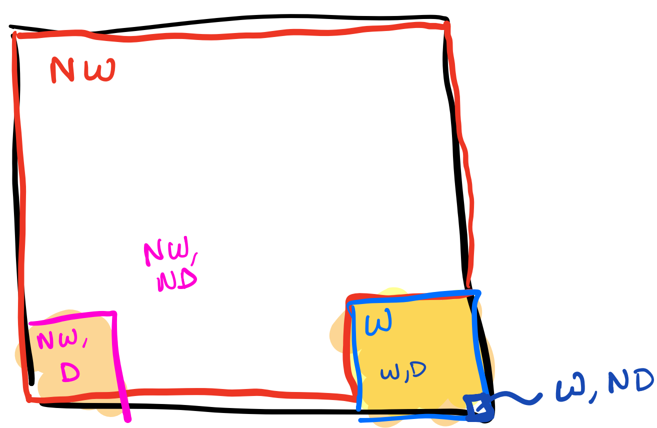
$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} \rightarrow \frac{P(A \cap B)}{P(A)}$$

1 quick Bayesian example : whales 

region: 95% there is no whale
5% there is a whale

device: correctly detect a whale 99% of the time
mistakenly detect whales 10% of the time

$$P = \begin{cases} 0.95 \text{ no whale} \begin{cases} 0.9 & \text{no detect} \\ 0.1 & \text{detect} \end{cases} \\ 0.05 \text{ whale} \begin{cases} 0.01 & \text{no detect} \\ 0.99 & \text{detect} \end{cases} \end{cases}$$



"if there is a detection, what is the prob. it was a whale"

$$P(W \mid D) = \frac{P(W \wedge D)}{P(D)} = \frac{P(W)P(W,D)}{P(W)P(W,D) + P(NW)P(NW,D)}$$

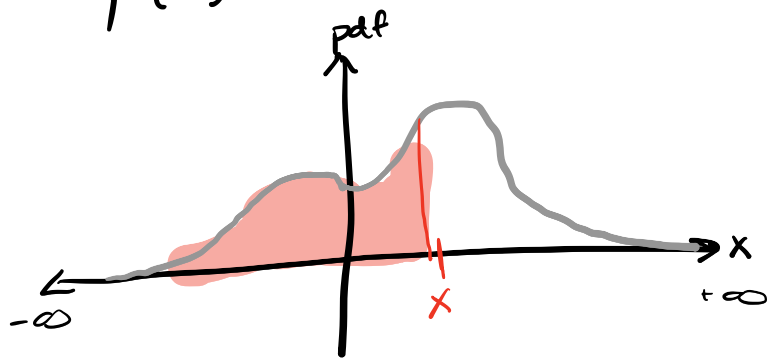
$$= \frac{(0.05)(0.99)}{(0.05)(0.99) + (0.95)(0.1)} = \boxed{0.34}$$

Bayes Thm: $P(A_i \mid B) = \dfrac{P(A_i)P(B \mid A_i)}{P(B)}$

STATISTICS

probability density function (PDF)

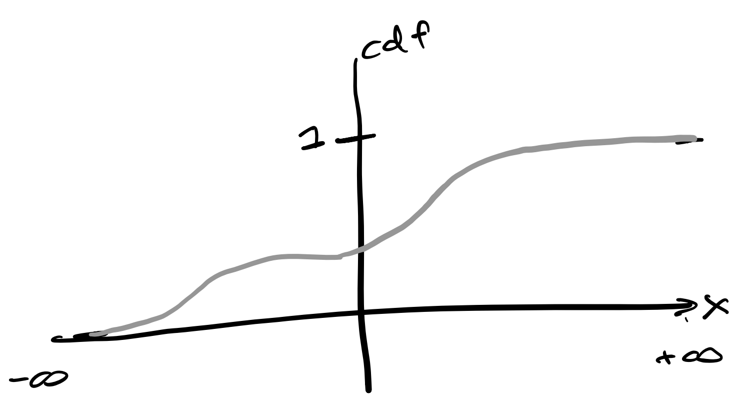
$p(x)$: "what's the probability of x



$$\int_{-\infty}^{+\infty} p(x) = 1$$

cumulative distribution function

$P_x(X)$: probability $x < X$



Ways To Understand 1 Distribution

- mean : $\frac{\sum_{i}^{N} x_i}{N} = \bar{x}$
- Variance: $\frac{\sum_{i}^{N} (x_i - \bar{x})^2}{N} = \sigma^2$

5 ± 0.3

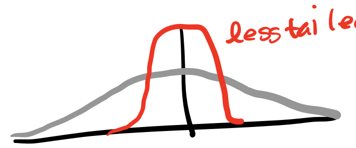
same units!

units of the variance are equal to the units of the measurement **squared**

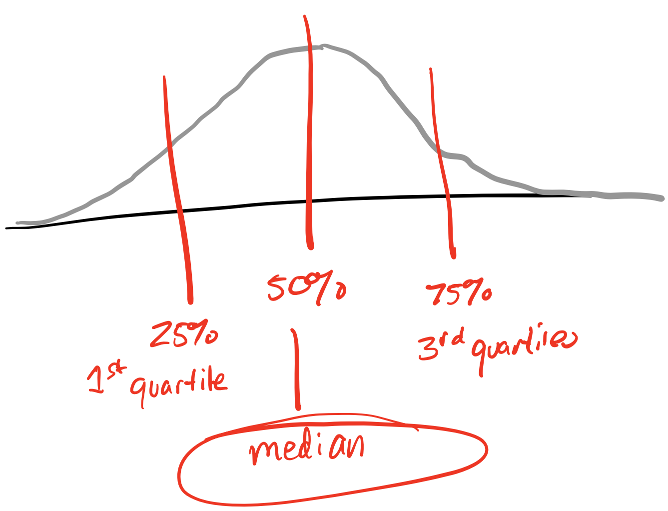
standard deviation: $\sqrt{\text{variance}}$

same units as measurements

- skew : left or right
- kurtosis : tailed



- quantiles : x% of the data/pdf occurs **below**



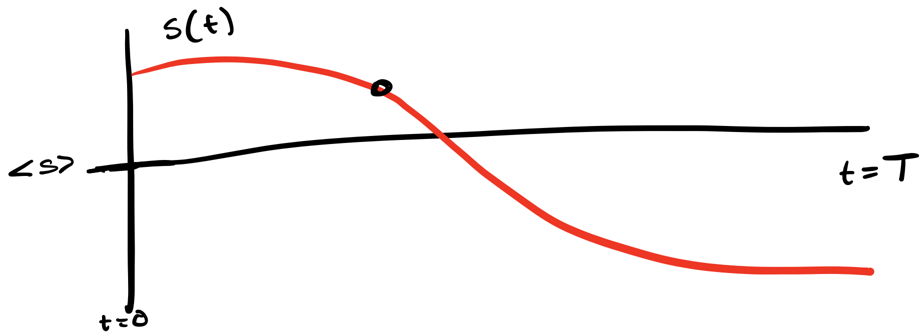
IQR : $(3^{rd}q - 1^{st}q)$

- mode (only for discrete)

Ways to Think about 2 Distributions

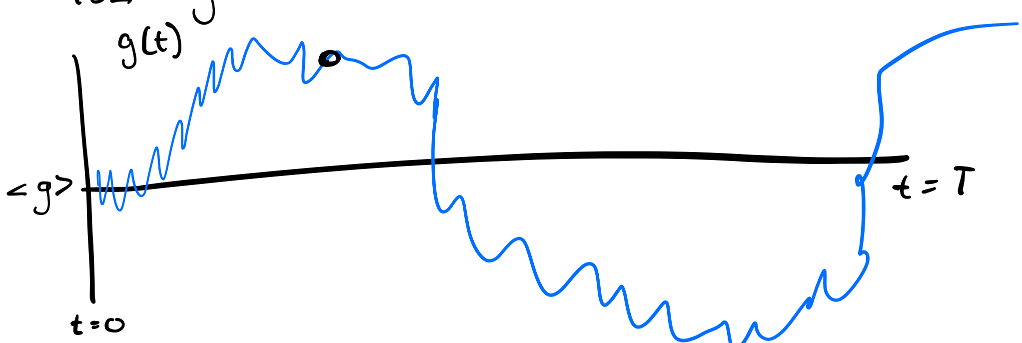
covariance : $\langle (X - \mu_x)(Y - \mu_y) \rangle$

TS1 : sun insolation



ω

TS2 : glaciation



m

$$\sum_{t=0}^{T} \frac{(s(t) - \langle s \rangle)(g(t) - \langle g \rangle)}{T}$$

$s(t)$	$>$ $\langle s \rangle$	AND $g(t)$	$>$ $\langle g \rangle$	: +
	$<$		$<$	: +
	$<$		$>$	: —
	$>$		$<$	: —

correlation : $\frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y}$ } -1 : 0 : 1

anticorr

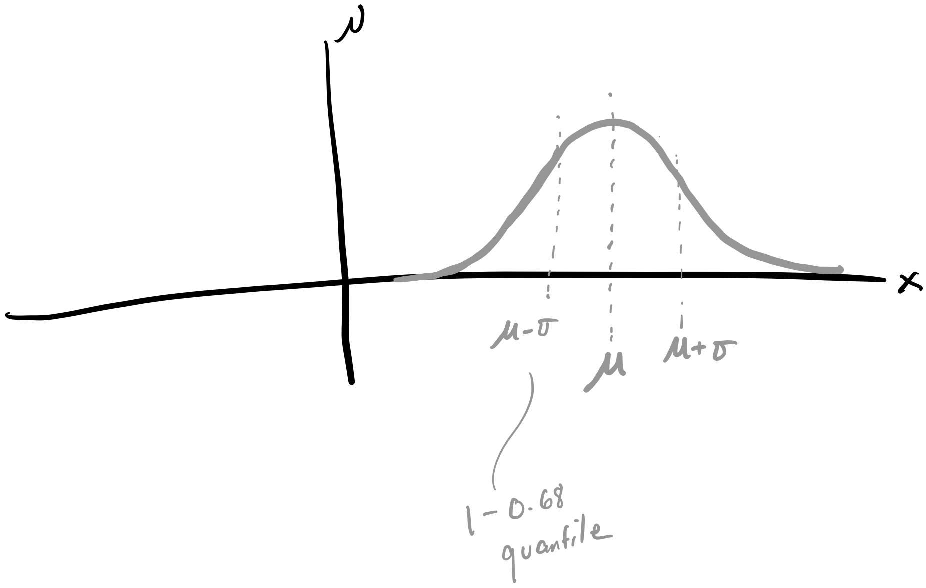
corr

not corr

Normal Distribution

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

(pdf)



lots of statistical properties assume normality

Central Limit Theorem : fluctuations between sample mean and the population mean are normally dist.

Hypothesis Testing

are your results significant

1) Form your null hypothesis

H_0 : "not significant" result

H_a : alternate hypothesis

H_0: " avg temp in TB = 850° " , H_a : " avg temp > 850° "

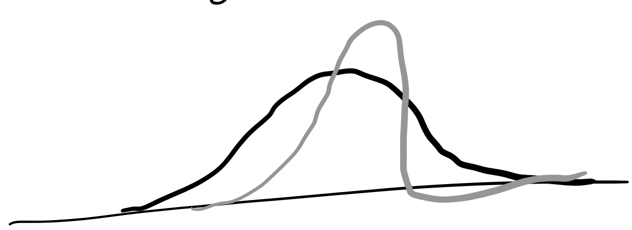
2) Determining an appropriate test statistic

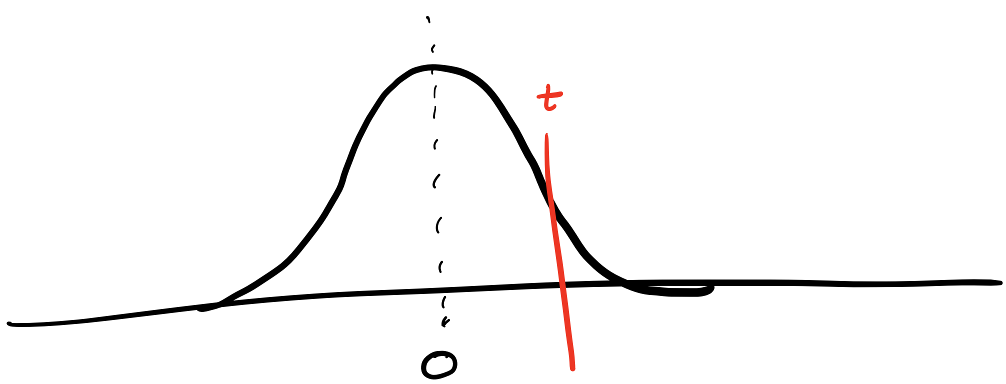
- $H_a: \mu > \mu_0$ (upper-tailed test)
- $H_a: \mu < \mu_0$ (lower-tailed test)
- $H_a: \mu \neq \mu_0$ (two-tailed test)

3) t-test : compute t-value

$$t = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

4) Map onto a $\underbrace{\text{Student's } t\text{-distribution}}_{\text{pdf: normal dist. scaled for d.o.f.}}$





compute p-value

probability of getting "t"
as a false positive

p-value < 0.05 → significant result

0.01

0.10 → reject H_0

p-value > 0.05 → insig.

fail to reject H_0