Data Mining: Probability and Statistics

Tom Heskes

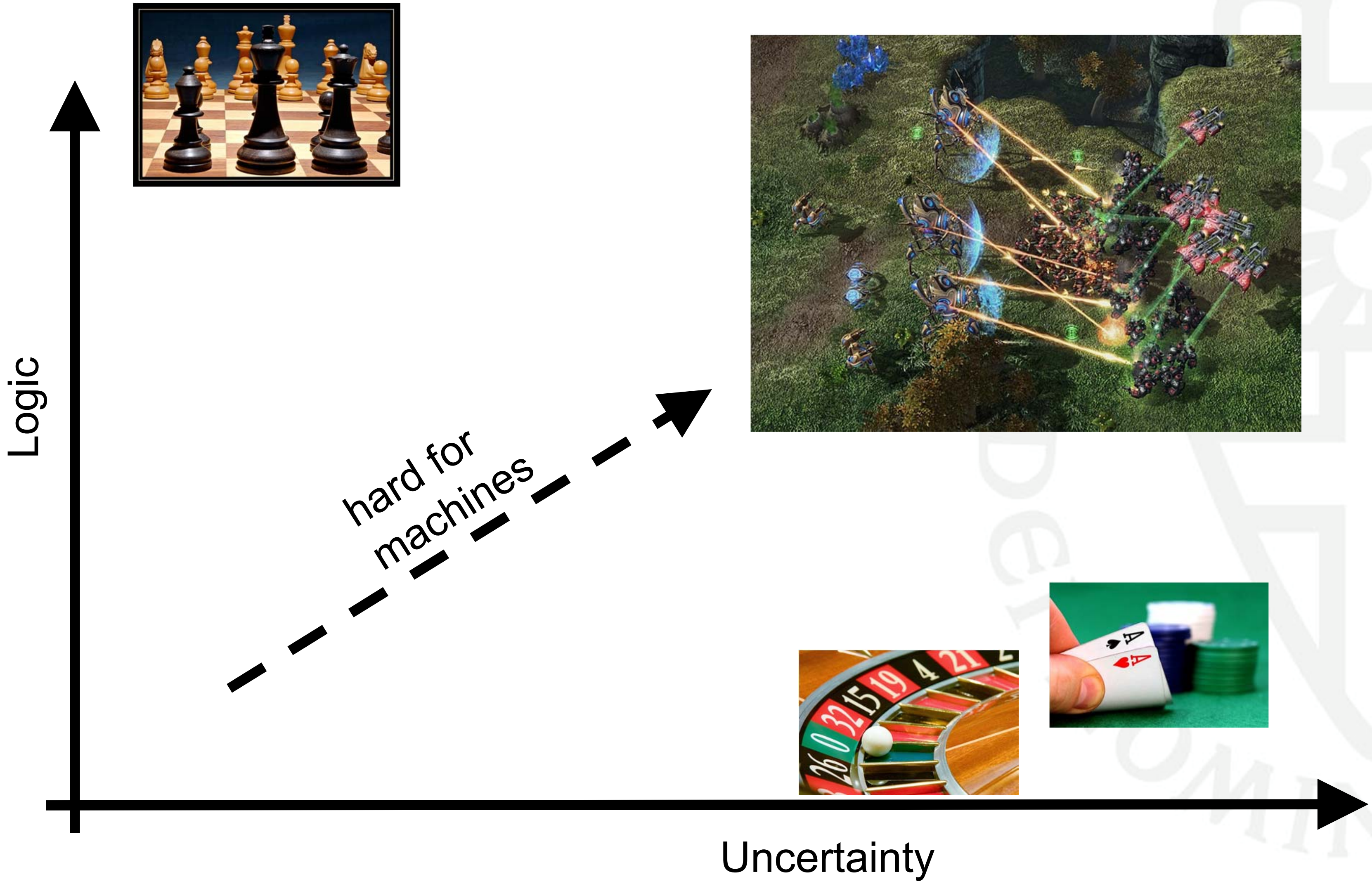
Radboud University Nijmegen

Probability and Statistics

- probability
- statistics
- hypothesis testing

- Note: see Appendix C of TSK

Probabilistic reasoning

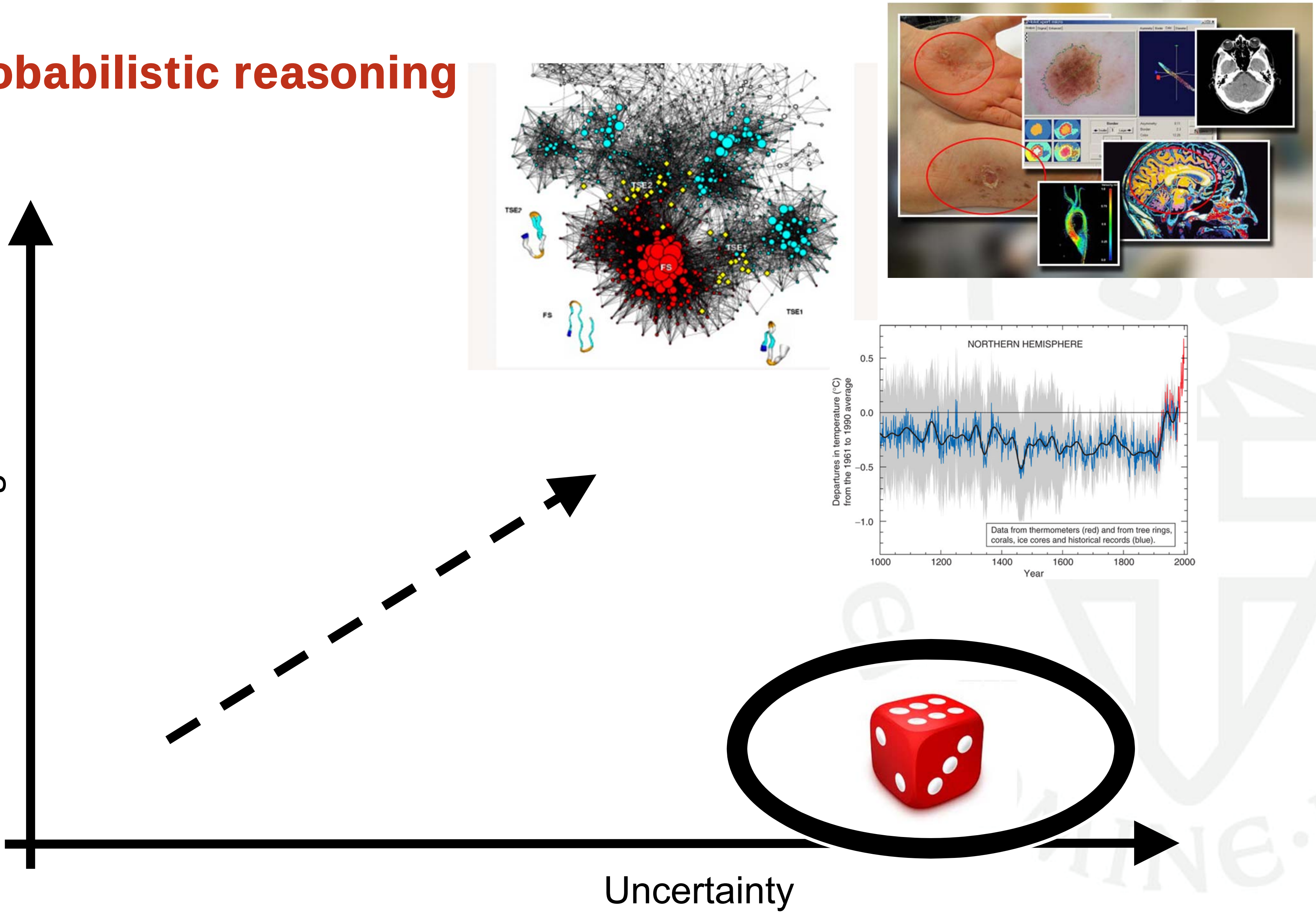


Logic

hard for machines

Uncertainty

Probabilistic reasoning



Concepts

- **Random experiment**
  - rolling a dice, flipping a coin, monitoring network traffic
- **Sample space**, all possible (single) outcomes: Ω
  - Ω = {1,2,3,4,5,6} for rolling a dice
  - Ω = {heads,tails} for flipping a coin
  - Ω = [0,+∞) for number of collisions per hour
- **Event** *E* is a subset of these outcomes: $E \subseteq \Omega$
  - *E* = {2,4,6} observing an even number

Probability

- A probability is a real-valued function define on the sample space Ω.
  - Probabilities are between 0 and 1:

$$E \subseteq \Omega : 0 \leq P(E) \leq 1$$

  - The probability of everything equals 1:

$$P(\Omega) = 1$$

  - Probabilities over disjoint events add:

$$\text{If } \; E_1 \cap E_2 = \varnothing \;\; \text{then} \;\; P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Random variable

- Quantity of interest related to a random experiment
  - number of heads when flipping a coin 30 times
  - time required to get back home
- Probability distribution (aka probability mass function) for a discrete random variable:

$$P(X = v) = P(E = \{e \mid e \in \Omega, X(e) = v\})$$

Probability distribution (example)

- A fair dice is rolled 4 times
- *X* is number of times the outcome is 3 or higher
- Possible outcomes: 6^4=1296
- Possible values for *X* are 0,1,2,3,4

X	0	1	2	3	4
P(*X*)	$(1/3)^4$ =1/81	$4(1/3)^3(2/3)$ =8/81	$6(1/3)^2(2/3)^2$ =24/81	$4(1/3)(2/3)^3$ =32/81	$(2/3)^4$ =16/81

Probability density function

- For continuous variables:

$$P(a < x < b) = \int_a^b f(x)\, dx$$

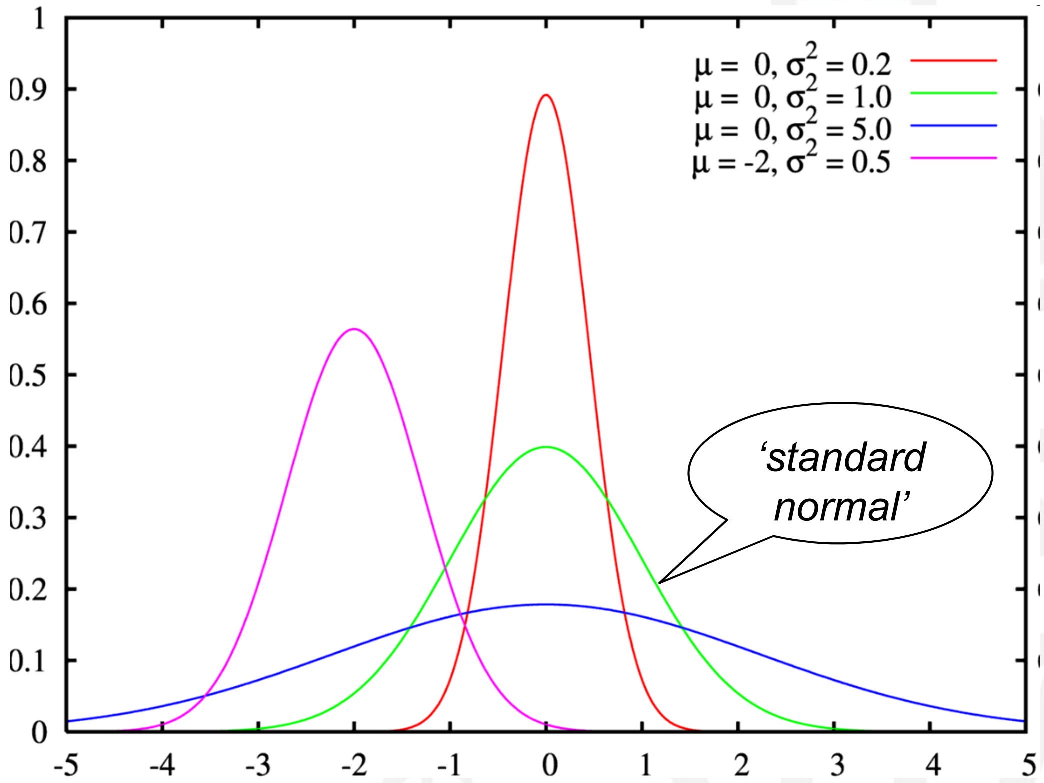
- *f(x)* is called a probability density function
- Probability that *X* takes a particular value is zero!
- Questions:
  - Can *f(x)* be negative?
  - Can *f(x)* be larger than 1?

Distribution plushies

Gaussian distribution

- Applicable in many fields due to **central limit theorem**
    - sum of many random variables is Gaussian
    - ‘error/noise model’

- Location parameter (**mean**) μ and spread (**standard deviation**) σ

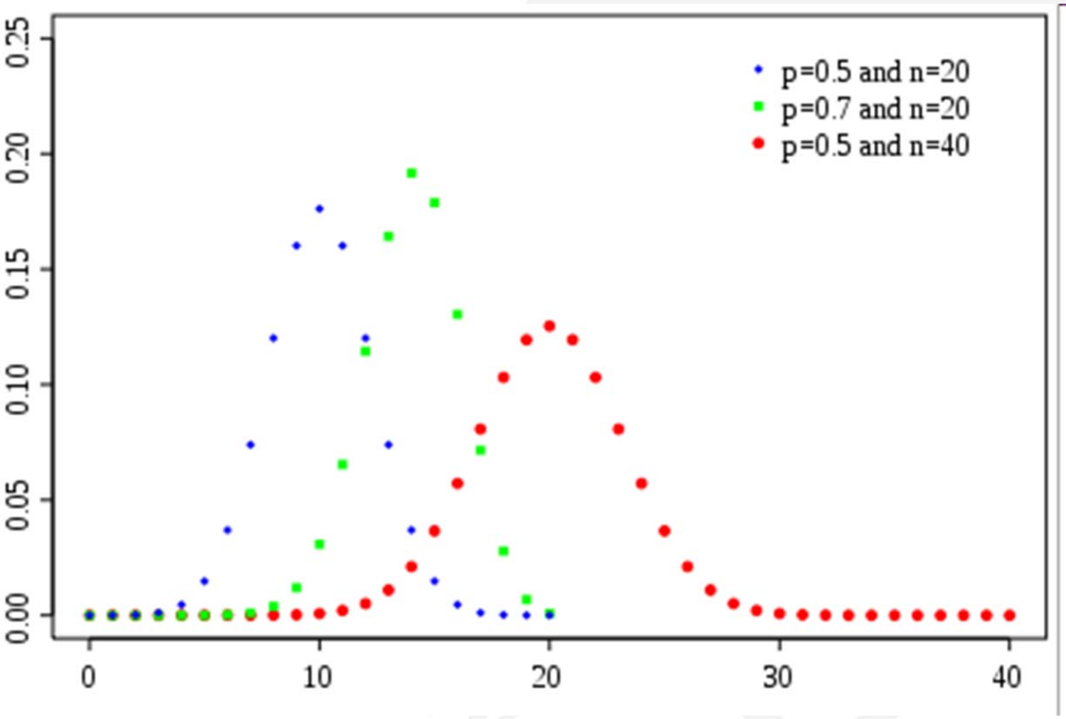


‘standard normal’

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Binomial distribution

- Number of successes in a number of independent yes/no trials
    - tossing a coin many times
    - nr. of ‘six-throws’ in a game of dice



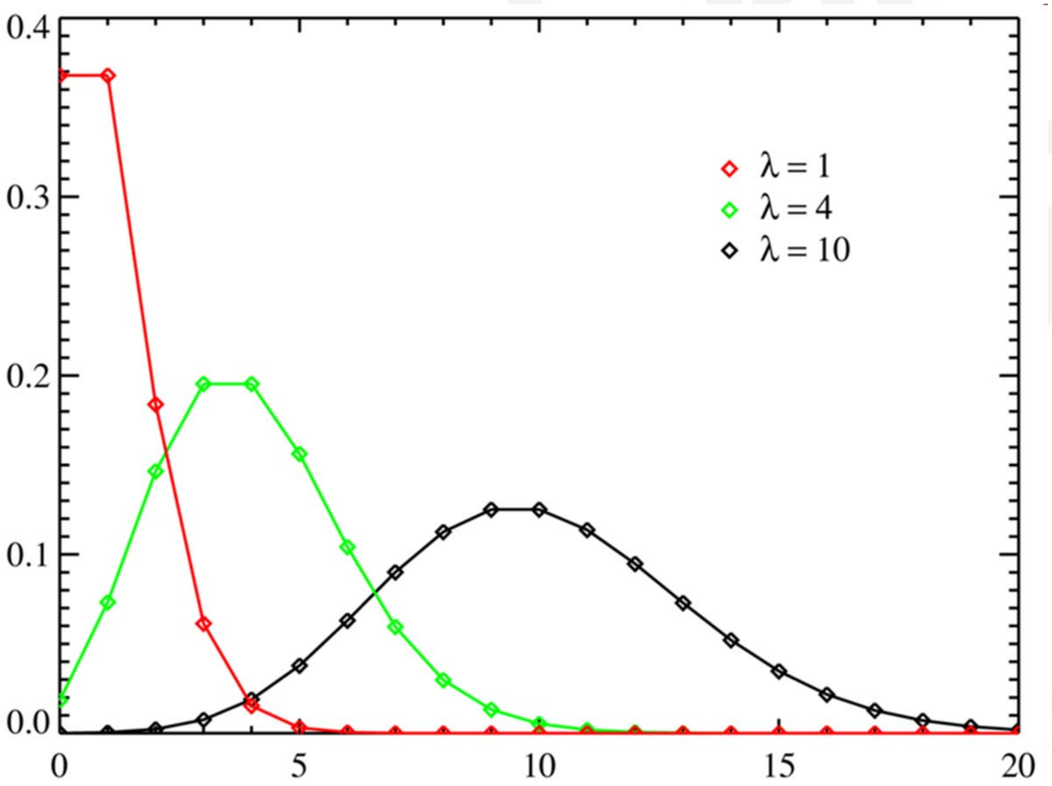
- Number of **trials** *n* and **probability of success** *p*

$$P(X = k; n, p) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Poisson distribution

- Probability of number of events occurring in a fixed period of time/space,
  - nr. of people entering the building per hour
  - nr. of hedgehogs killed per km of road
  - nr. of mutations per 100.000 base pairs
- typically ‘rare events’
- **Rate** parameter *λ*

$$P(X = k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$



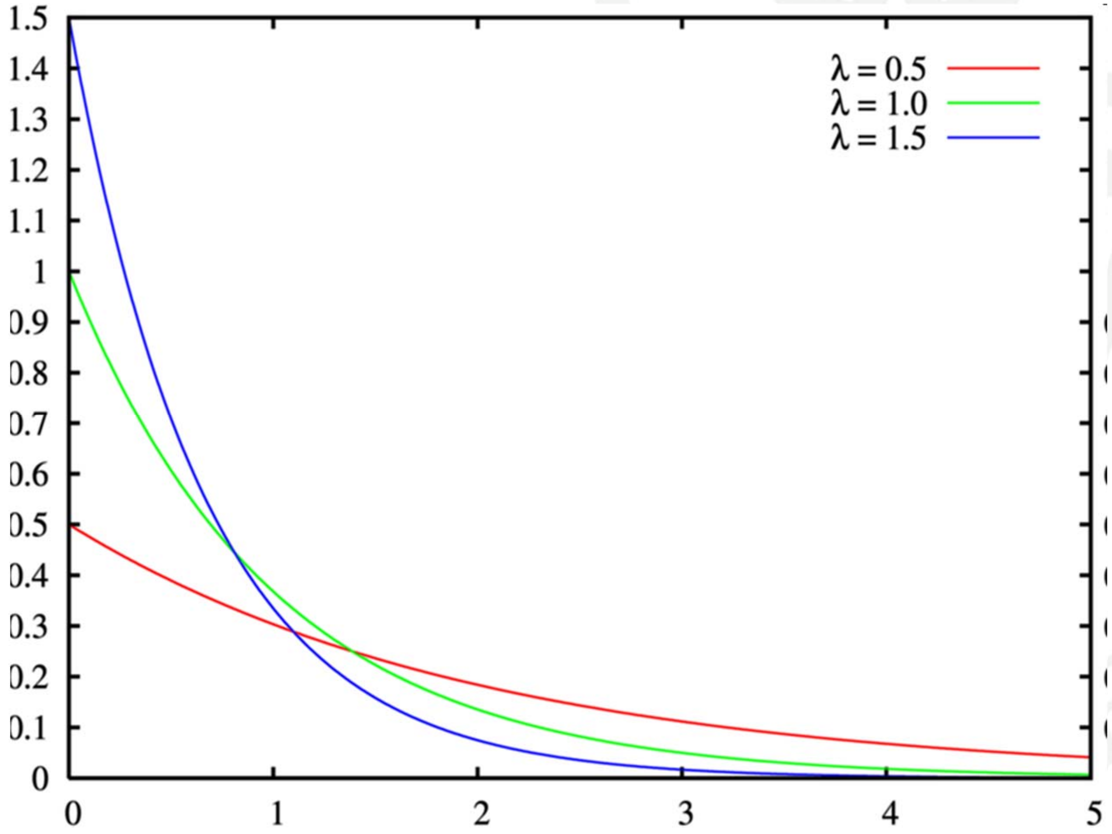
Exponential distribution

- Probability density of times *between* events, e.g.,
  - time it takes before the next person enters the building
  - time between hits on a website

- ‘Memoryless’

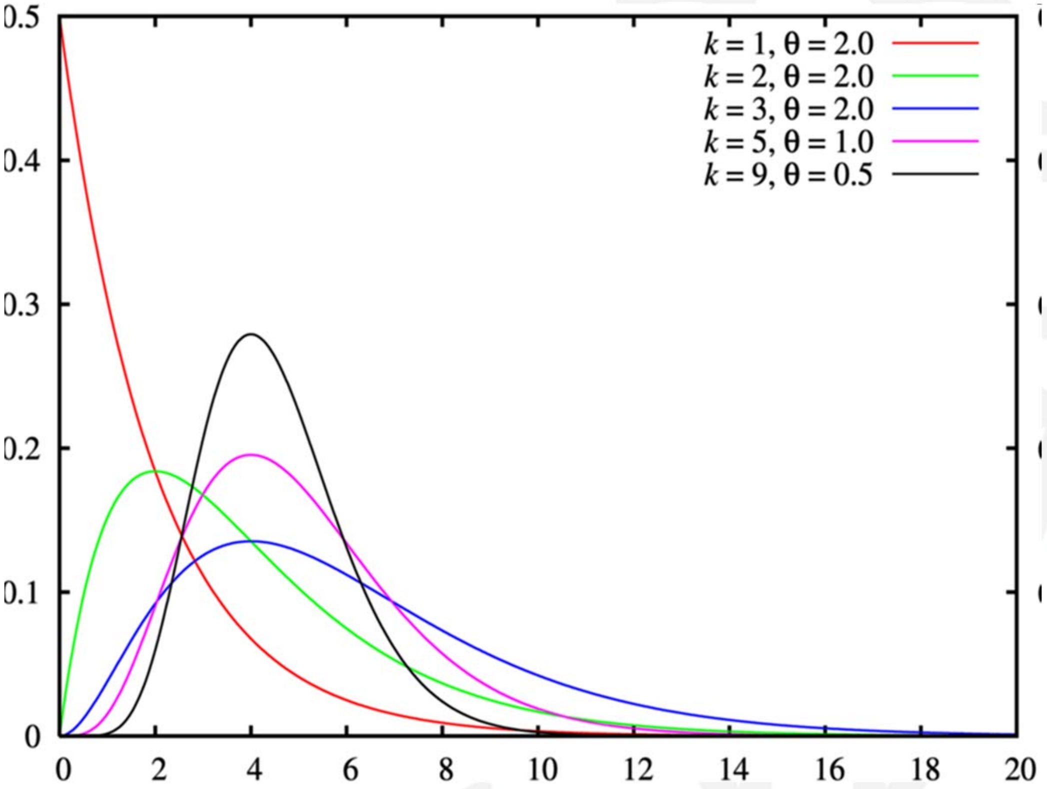
- **Rate** parameter λ

$$f(x;\lambda) = \lambda\, e^{-\lambda x}$$



Gamma distribution

- “Gaussian” for only positive values,
  - distribution of incomes
  - lifetime of light bulbs

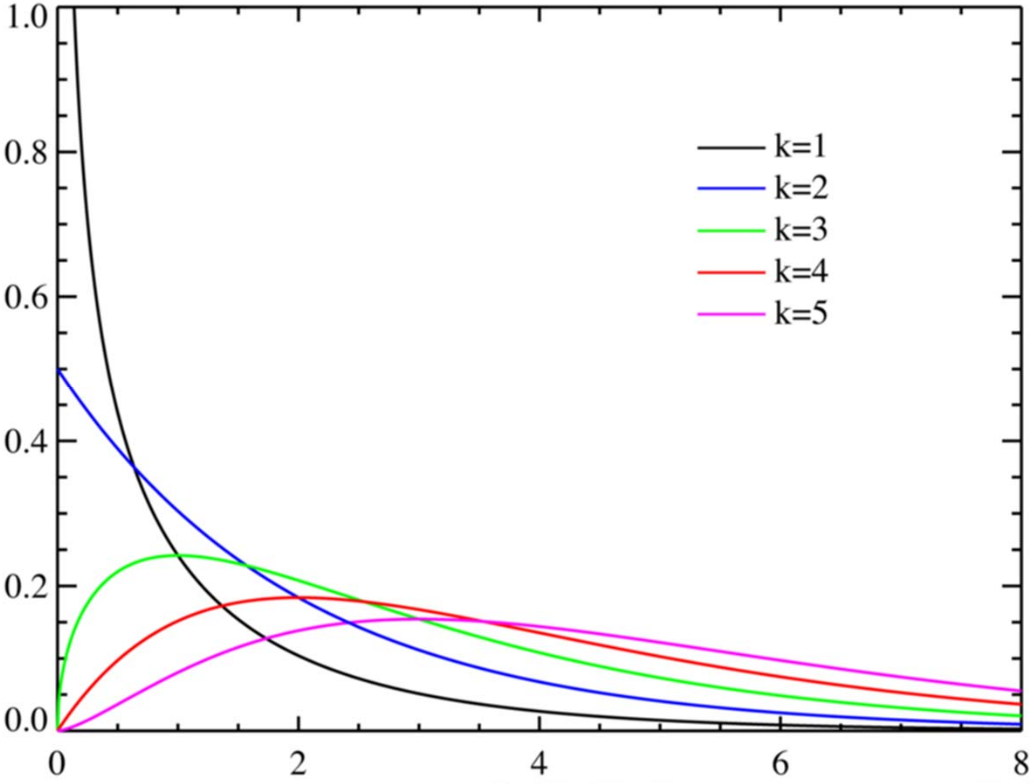


- **Scale** parameter θ and **shape** parameter k

$$f(x;\theta,k) = \frac{x^{k-1}\; e^{-x/\theta}}{\theta^k \Gamma(k)}$$

Chi-square distribution

- Often used in statistical significance tests
- Special case of Gamma distribution (with $\theta \to 2$, $k \to k/2$)

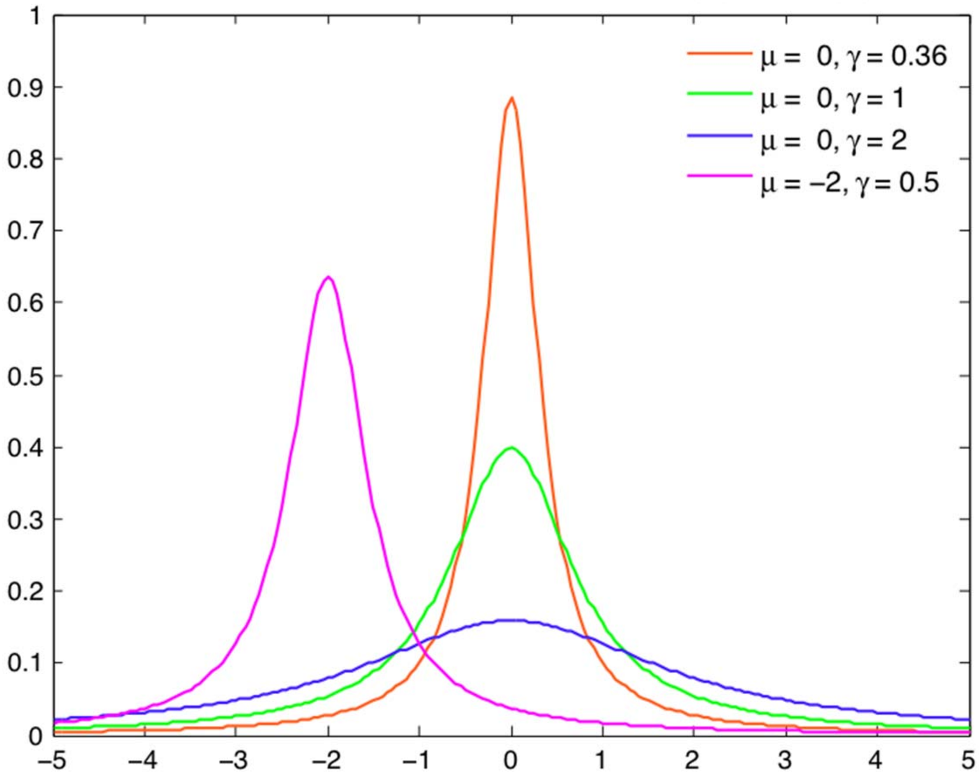


- **Degrees of freedom** *k*: (distribution of sum of the squares of *k* normally distributed random variables)

$$f(x;k) = \frac{x^{(k/2)-1}\, e^{-x/2}}{2^{k/2}\Gamma(k/2)}$$

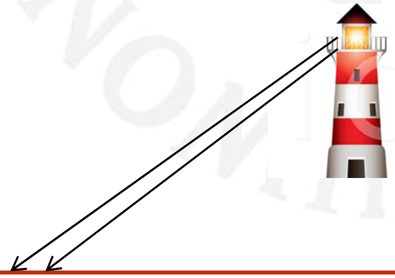
Tails matter …

- Cauchy distribution
  - looks like a ‘fat tailed’ Gaussian …
  - … but has no mean(!), no variance
  - and very insensitive to outliers



- Location parameter *μ* and scale parameter *γ*

$$f(x;\mu,\gamma)=\frac{1}{\pi\gamma\left[1+\left(\frac{x-\mu}{\gamma}\right)^{2}\right]}$$



Multiple random variables

- If X and Y are two random variables, then $P(X, Y)$ is their joint probability distribution
- If the random variables are **independent**, we have

$$P(X,Y) = P(X)P(Y)$$

- Example: Throwing a fair dice
  - X: outcome of die is '3' or higher;
  - Y: even outcome

⇒ Are X and Y independent?

- $P(X) = P(\{3,4,5,6\}) = 2/3$,
- $P(Y) = P(\{2,4,6\}) =1/2$,
- $P(X,Y) = P(\{4,6\}) = 1/3 = P(X)\ P(Y)$, …. so yes: independent

Conditional probability

- Definition:

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)}$$

- Probability of *Y* "given" *X*

- Example: Throwing a fair dice
  - *X*: outcome of die is '3' or higher;
  - *Y*: even outcome

  ⇒ What is $P(Y|X)$?

  - direct: $P(Y|X) = P(\{4,6\} \mid \{3,4,5,6\}) = \frac{1}{2}$
  - formula: $P(Y|X) = (P(X,Y) = \frac{1}{3}) / (P(X) = \frac{2}{3}) = \frac{1}{2}$

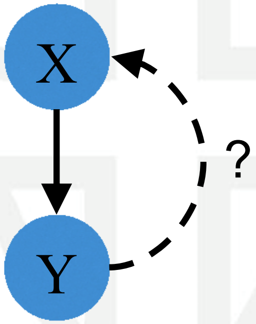
Bayes' theorem

- From $P(Y \mid X) = \dfrac{P(X,Y)}{P(X)}$

  and $P(X \mid Y) = \dfrac{P(X,Y)}{P(Y)}$

- we have

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$



- Using Bayes' rule we can **invert** the probability of *effect given cause* to the probability of *cause given effect*: **probabilistic reasoning**

Expected value (discrete)

- The **expected value** of a function *g* of a discrete random variable *X*:

$$E[g(X)] = \sum_k g(k) P(X = k)$$

- Example:
  - If you throw outcome *k*, you receive k^2 euros
  - What is your expected pay-off for a fair dice?
  -

$$E[k^2] = \sum_{k=1}^{6} k^2 \frac{1}{6} = \frac{1+4+9+16+25+36}{6} = \frac{91}{6}$$

Expected value (continuous)

- The expected value of a function *g* of a continuous random variable *X*:

$$E[g(X)] = \int g(x) f(x)\, dx$$

- Example:
  - *X* homogeneously distributed between 0 and 1
  - What is $E[x^2]$?
  -

$$E[x^2] = \int_0^1 x^2 1\, dx = \frac{1}{3} x^3 \Big|_0^1 = \frac{1}{3}$$

Common expected values

- Mean value:

$$\mu_X = E[X] = \sum_k k\ P(X = k) \text{ or } \mu_X = \int x\ f(x)\,dx$$

- Variance:

$$\sigma_X^2 = Var[X] = E[(X - \mu_X)^2] = E[X^2] - \mu_X^2$$

- Covariance:

$$Cov[X, Y] = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y$$

Expectation 'paradoxes'

- Devil's wager:
  - you are accidentally and unexpectedly sent to hell, where Old Harry makes you an offer:
  - toss a fair coin → heads you're free, tails you stay … forever, or
  - wait one day in hell → then get the same offer but with half the chance to loose the bet

$$E[tomorrow] = -1000 + \frac{p}{2}(-Inf) + \left(1 - \frac{p}{2}\right) \cdot 0 > p(-Inf) = E[today]$$

  - … so the optimal solution is to never leave?

- Two envelopes problem:
  - one envelop contains twice as much money as the other, you get to choose
  - you want to pick A … but what if you switch to B?

$$E[B] = \sum_k k\, P(X = k) = \frac{1}{2}(2A) + \frac{1}{2}\left(\frac{A}{2}\right) = \frac{5}{4}A \quad > E[A]$$

  - … and back again? And again?

Statistics

- “Inverse” probability theory

- **Probability**: given the rules of probability theory, compute probabilities and expected values of interest given a particular probability model

- **Statistics**: given a finite set of data (and assuming some underlying probability model), estimate the parameters of the model

Point estimation

- **Model**: N samples X_i are drawn from some probability density with (unknown) mean μ_X and variance σ_X^2
- Given data, what's our best estimate for μ_X and σ_X^2?
- Obvious choices:
  - sample mean

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

  - sample variance

$$s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{x}\right)^2$$

Unbiased estimator

- Thought experiment: repeat the previous many times, i.e.
  - generate *N* samples X_i from some probability density with mean μ_X and variance σ_X^2
  - compute the resulting **sample mean** and **sample variance**
  - Check whether, **on average**, the answer is correct

- Easy to check for the sample mean:

$$E\left[\overline{X}\right] = E\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N}\sum_{i=1}^{N} E\left[X_i\right] = \frac{1}{N}\sum_{i=1}^{N} \mu_X = \mu_X$$

Sample variance (1)

$$E\left[S_X^2\right] = E\left[\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2\right] = \frac{1}{N-1}E\left[\sum_{i=1}^{N}\left(X_i - \frac{1}{N}\sum_{j=1}^{N}X_j\right)^2\right]$$

$$= \frac{1}{N-1}E\left[\sum_{i=1}^{N}\left(X_i^2 - \frac{2}{N}\sum_{j=1}^{N}X_iX_j + \left\{\frac{1}{N}\sum_{j=1}^{N}X_j\right\}^2\right)\right]$$

$$= \frac{1}{N-1}E\left[\sum_{i=1}^{N}X_i^2 - \frac{2}{N}\sum_{i,j=1}^{N}X_iX_j + \frac{1}{N}\left\{\sum_{j=1}^{N}X_j\right\}^2\right]$$

$$= \frac{1}{N-1}E\left[\sum_{i=1}^{N}X_i^2 - \frac{1}{N}\sum_{i,j=1}^{N}X_iX_j\right] = \frac{1}{N-1}E\left[\sum_{i=1}^{N}X_i^2 \overbrace{- \frac{1}{N}\sum_{i=1}^{N}X_i^2 - \frac{1}{N}\sum_{i,j=1;\,j\neq i}^{N}X_iX_j}^{\text{this is where it happens…}}\right]$$

Sample variance (2)

- From previous slide:

$$E\left[S_X^2\right] = \frac{1}{N-1} E\left[\sum_{i=1}^{N} X_i^2 - \frac{1}{N}\sum_{i=1}^{N} X_i^2 - \frac{1}{N}\sum_{i,j=1;j\neq i}^{N} X_i X_j\right]$$

- From definitions and independent samples:

$$\sigma_X^2 = E[X^2] - \mu_X^2$$

$$Cov[X,Y] = E[XY] - \mu_X \mu_Y$$

$$E\left[X_i^2\right] = \mu_X^2 + \sigma_X^2; \quad E\left[X_i X_j\right] = \mu_X^2 \text{ if } j \neq i$$

- And thus:

$$E\left[S_X^2\right] = \frac{1}{N-1}\left[N\left(\mu_X^2 + \sigma_X^2\right) - \frac{1}{N} N\left(\mu_X^2 + \sigma_X^2\right) + \frac{1}{N} N(N-1)\mu_X^2\right] =$$

$$= \frac{1}{N-1}\left[(N-1)\left(\mu_X^2 + \sigma_X^2\right) - (N-1)\mu_X^2\right] = \sigma_X^2$$

Standard error of the mean

- Using similar calculations, it can be shown that

$$E\left[\left(\overline{X} - \mu_X\right)^2\right] = \frac{1}{N}\sigma_X^2$$

- Substitute the estimate s_X for the (unknown) σ_X

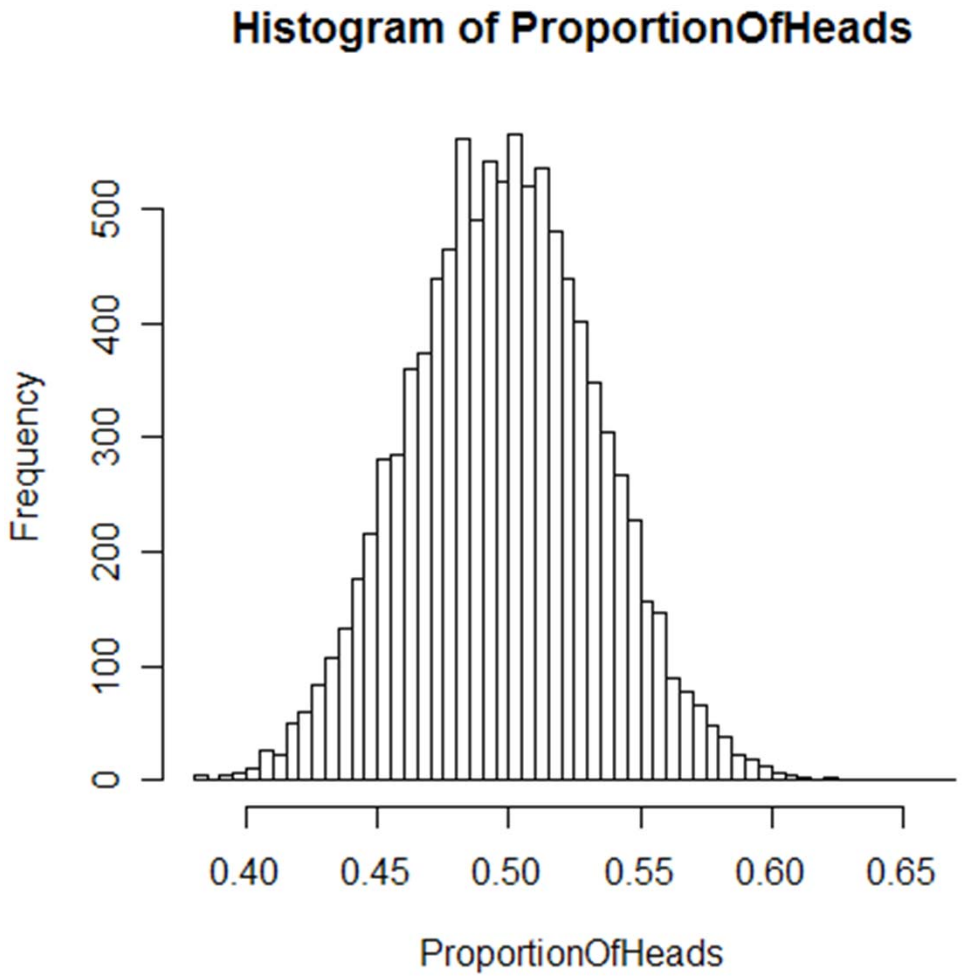
- $s_X / \sqrt{N}$ is called the *standard error of the mean*

Central limit theorem

- Consider the sample mean $\overline{X}$ of N samples from some distribution with **mean** μ_X and **variance** σ_X^2

- For large N, the distribution of the sample mean $\overline{X}$ approaches a **Gaussian** with mean μ_X and variance σ_X^2/N

- This is **independent** of the underlying distribution of the samples!



Interval estimation

- We'd like to say a bit more than just our best guess
- Next best: mention the **standard error**
- Even better: give a **confidence interval**

$$P(\theta_1 < \theta < \theta_2) = 1 - \alpha$$

- (θ_1, θ_2) is the confidence interval for θ at the **confidence level** α

Interpretations of confidence interval

- “Were this procedure to be repeated on multiple samples, the calculated confidence interval (which would differ for each sample) would encompass the true population parameter 90% of the time”

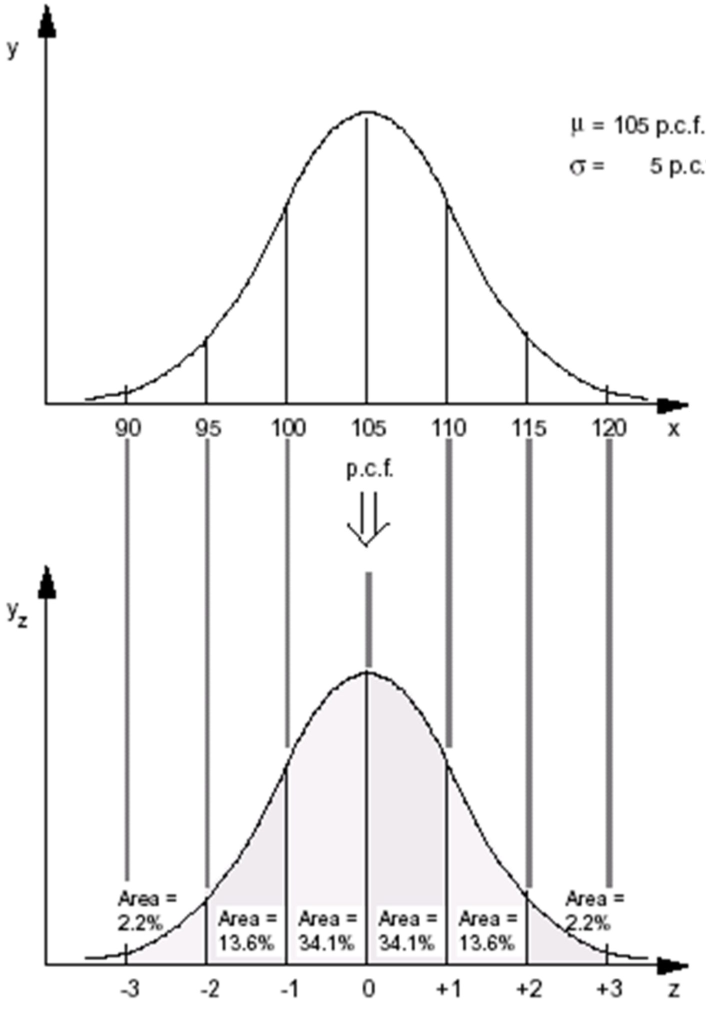
- “The confidence interval for α=0.1 represents values for the population parameter for which the difference between the parameter and the observed estimate is not statistically significant at the 10% level”

Confidence interval for sample mean (1)

- **Central limit theorem**: the distribution of the population mean $\bar{X}$ approaches a normal distribution with mean μ_X and variance σ_X^2/N
- That is, the variable $$Z = \frac{\bar{X} - \mu_X}{\sigma_X / \sqrt{N}}$$

  has a **standard normal** distribution (mean 0, variance 1):

$$P(\mu_X - z^* \sigma_X / \sqrt{N} < \bar{X} < \mu_X + z^* \sigma_X / \sqrt{N})$$
$$= P(-z^* < Z < z^*)$$



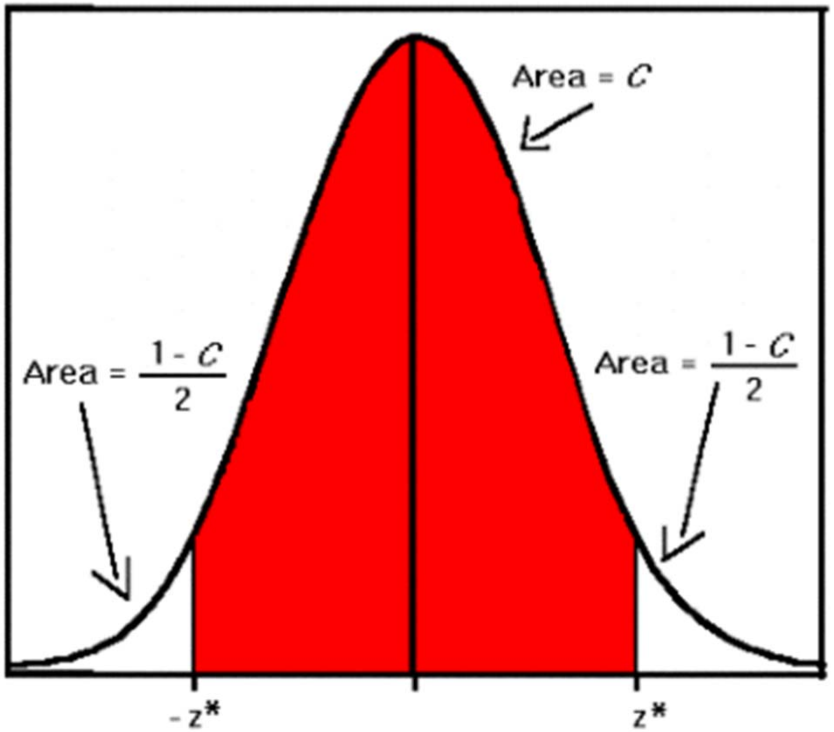
Confidence interval for sample mean (2)

- "Inverting" this, if we **observe** a sample mean $\overline{X}$, the confidence interval for μ_X reads

$$P(\overline{x} - z^* \sigma_X / \sqrt{N} < \mu_X < \overline{x} + z^* \sigma_X / \sqrt{N}) = P(-z^* < Z < z^*)$$

- We typically don't know σ_X and then substitute our **best estimate** s_X

$$P(\overline{x} - z^* s_X / \sqrt{N} < \mu_X < \overline{x} + z^* s_X / \sqrt{N})$$
$$= P(-z^* < Z < z^*)$$



Hypothesis testing

- Should we **accept** or **reject** a hypothesis (e.g., ‘men are taller than women’) given the data available?
- Typical question in data mining: is one method or model *significantly* better than another?
- Results are often only publishable if they show a significant improvement at significance level α=0.05

Confirmatory data analysis

- Assuming that the **null hypothesis** is **true**, what is the **probability** of observing a value for the **test statistic** that is **at least as extreme** as the value that was actually observed?

- Null hypothesis:
  - coin/dice is fair,
  - no difference between classification methods,
  - random variables X and Y are independent, …

- Test statistic:
  - number of heads,
  - difference between performance scores,
  - chi-squared statistic as normalized sum of squared difference between observed and expected frequencies under the null hypothesis, …

Procedure

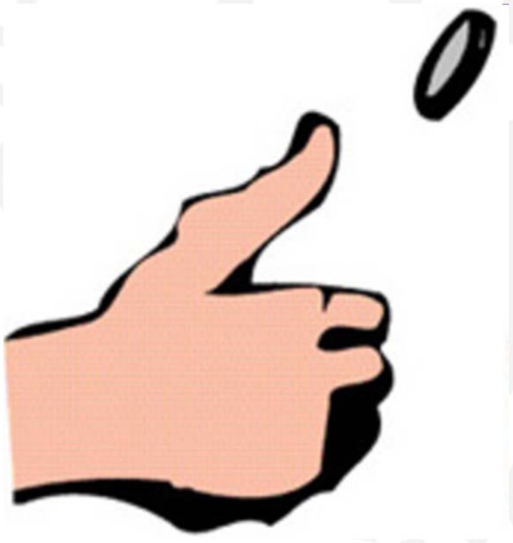
- Formulate the **null** (“simple”) **hypothesis**
- Define a **significance level** α
- Define a **test statistic** θ with a known probability distribution under the null hypothesis
- Compute θ^* as the value of θ from the **observed data**
- Compute the ***p*-value**: the probability of θ under the null hypothesis at least as extreme as the observed value θ^*
- **Reject** the null hypothesis if the *p*-value is **smaller** than the significance level α

In terms of confidence intervals

- Formulate the *null* (“simple”) *hypothesis*
- Define a *significance level* α
- Define a *test statistic* θ with a known probability distribution under the null hypothesis
- Compute the value of θ from the observed data
- Compute the **confidence interval** for θ under the null hypothesis for confidence level α
- **Reject** the null hypothesis if the observed value θ^* is **outside** the confidence interval

Example: fair coin (1)

- Null hypothesis: our coin is fair
- Choose significance level, e.g. α=0.05
- Observed data: N=100 throws, 60 heads, 40 tails
- Enough evidence to reject the null hypothesis?



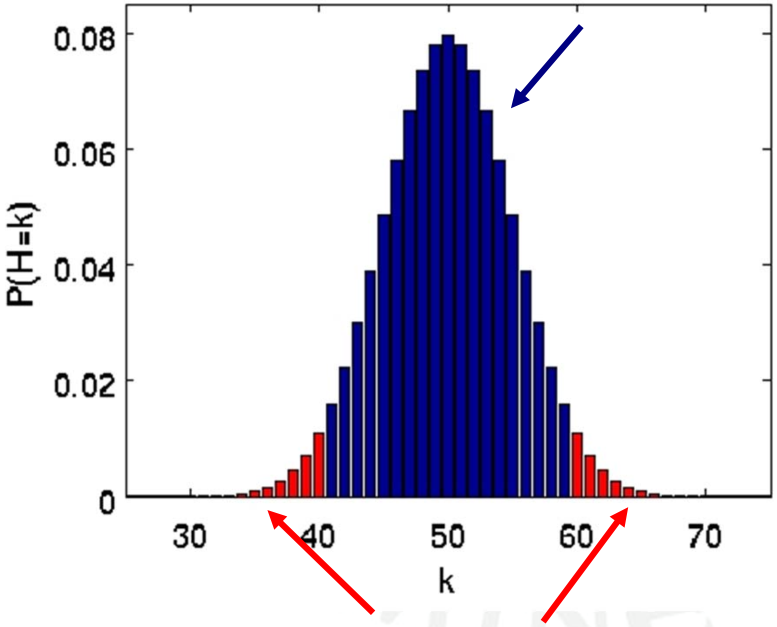
Example: fair coin (2)

- Test statistic: *H* = number of heads
- Observed: *H**=60
- Probability distribution of *H* under null hypothesis: binomial distribution

$$P(H = k) = \binom{N}{k} 0.5^k \,(1-0.5)^{N-k} = \binom{N}{k} 0.5^N$$

- *p*-value (red area): 0.057, i.e., *not significant* at 0.05 level: no (not enough) reason to reject the null hypothesis

less extreme than observed



at least as extreme as observed

One-sided versus two-sided tests

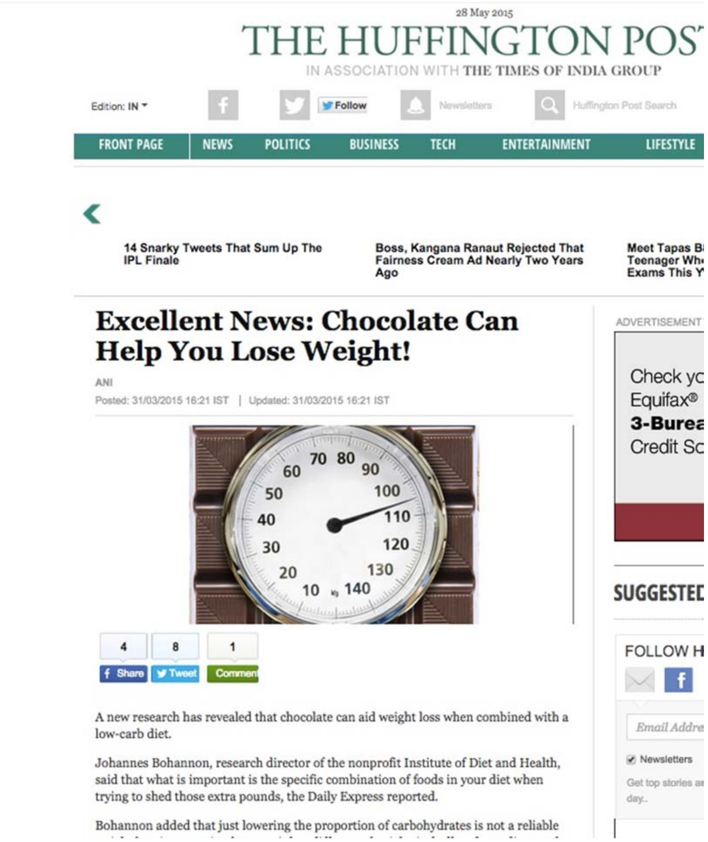
- One-sided:
  - “better/larger/heavier than”
  - consider only one of the tails to compute p-value

- Two-sided:
  - “different from”
  - consider both tails to compute p-value
  - (or consider one tail, but then divide the significance level by 2)

Publication bias and p-value hunting

- Results that are not statistically significant are still hard to publish…
- Publication bias
- P-value hunting

Chocolate accelerates weight loss



Excellent News: Chocolate Can Help You Lose Weight!

ANI

Posted: 31/03/2015 16:21 IST | Updated: 31/03/2015 16:21 IST

A new research has revealed that chocolate can aid weight loss when combined with a low-carb diet.

Johannes Bohannon, research director of the nonprofit Institute of Diet and Health, said that what is important is the specific combination of foods in your diet when trying to shed those extra pounds, the Daily Express reported.

Bohannon added that just lowering the proportion of carbohydrates is not a reliable

---

Weight Loss FOODS FOR WEIGHT LOSS

Lose 10% More Weight By Eating A Every Day...No Joke

APRIL 24, 2015 By AVIVA PATZ

WRITE A COMMENT

Want to lose weight faster? Eat dark chocolate. That's right, the same delicious bar that we know is packed with healthy antioxidants just keeps getting better: It's now a promising weight-loss aid, according to findings published last week in the *International Archives of Medicine*.

Dark chocolate has enjoyed a health halo for years now, thanks to its rich

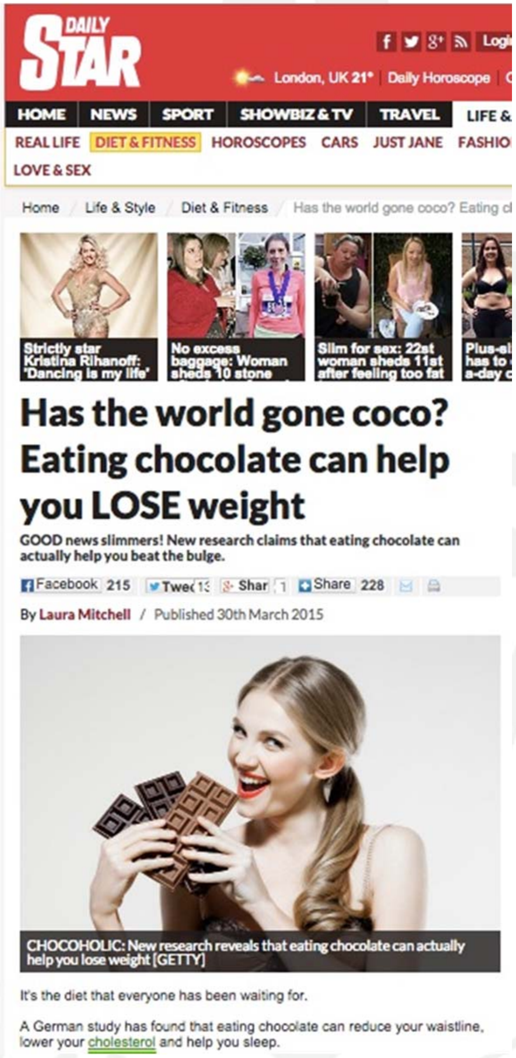
---

03.04 02:35 MIGnews.com

Шоколад - лучшая диета

Сотрудники немецкого Института питания и здоровья провели исследование, в результате которого пришли к выводу, что шоколад в сочетании с низкоуглеводной диетой помогает быстрее похудеть.

В ходе эксперимента его участники 19-67 лет были разделены на три группы. Первая группа соблюдала низкоуглеводную диету, вторая помимо диеты употребляла по 42 грамма темного шоколада в день, в



Has the world gone coco? Eating chocolate can help you LOSE weight

GOOD news slimmers! New research claims that eating chocolate can actually help you beat the bulge.

By Laura Mitchell / Published 30th March 2015

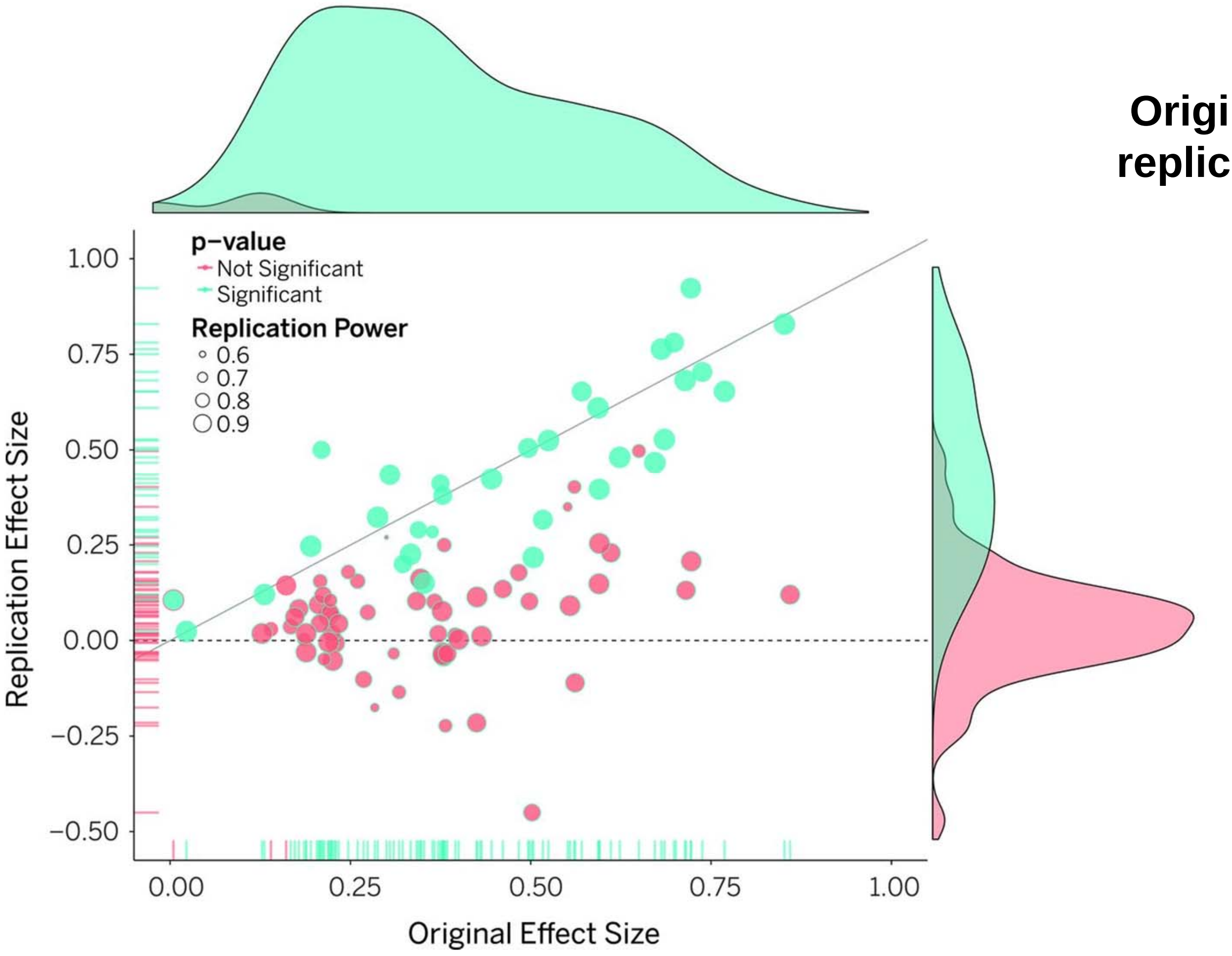
CHOCOHOLIC: New research reveals that eating chocolate can actually help you lose weight [GETTY]

It's the diet that everyone has been waiting for.

A German study has found that eating chocolate can reduce your waistline, lower your cholesterol and help you sleep.

http://io9.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800

Reproducibility of psychological science

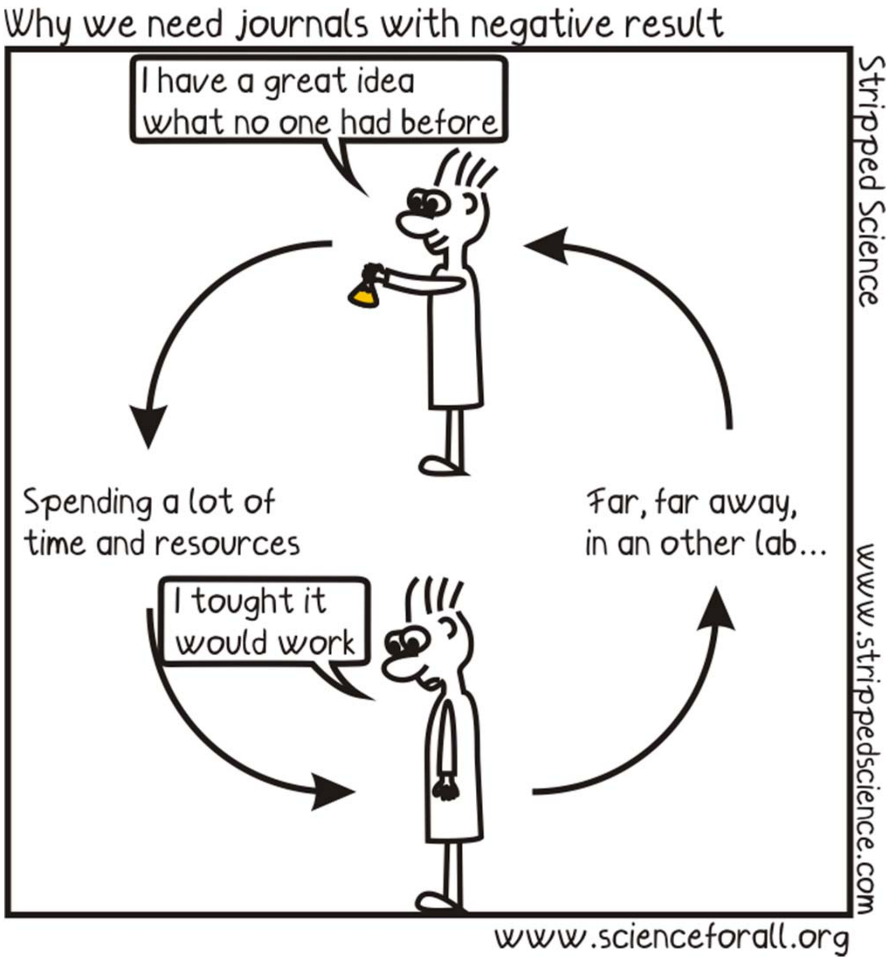


Original study effect size versus replication effect size (correlation coefficients).

Open Science Collaboration Science 2015;349:aac4716

Published by AAAS

Publishing negative results



The Null Journal

Negative Results, Failures, and Voids

JOURNAL OF NEGATIVE RESULTS IN BIOMEDICINE