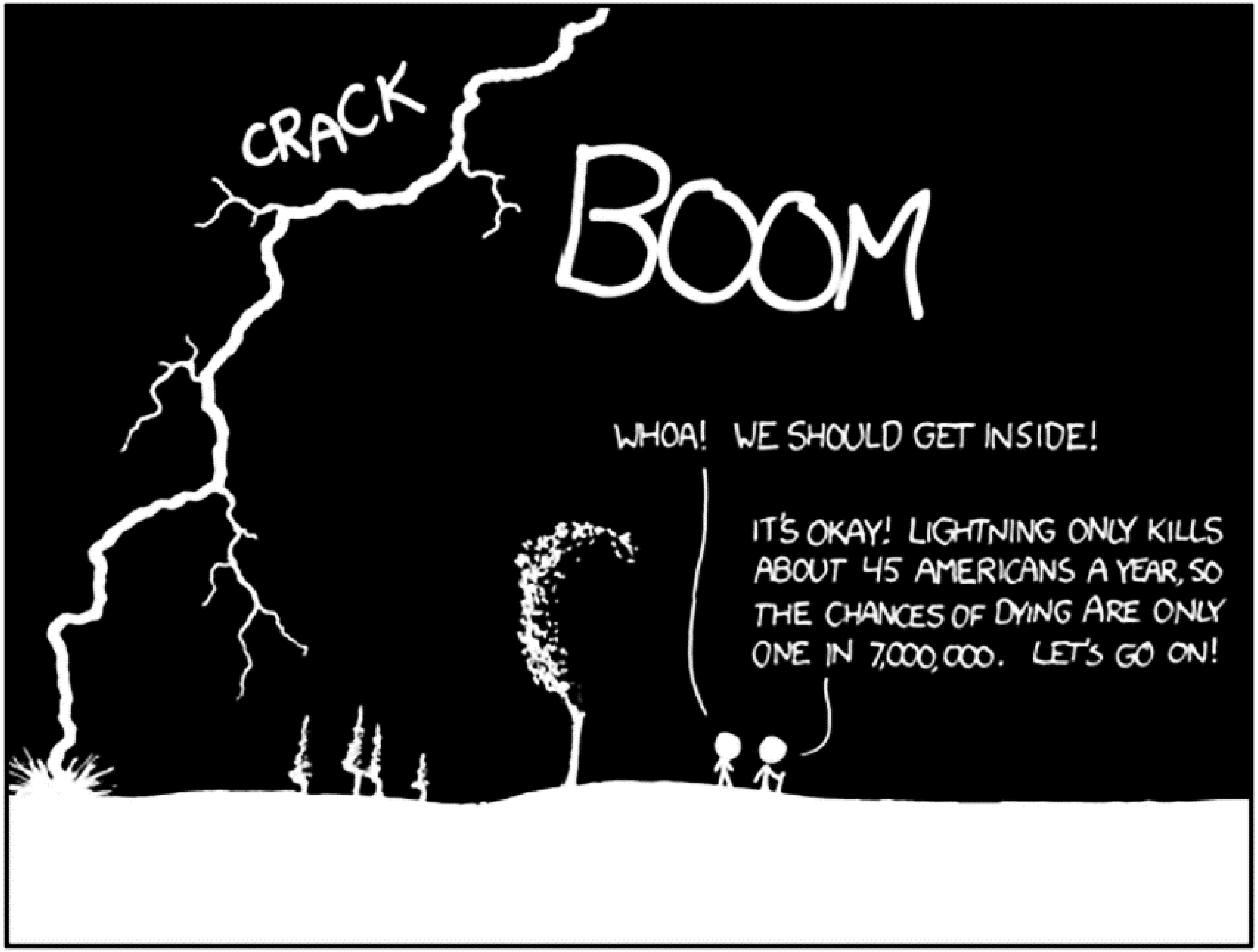


THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

Probability and Statistics

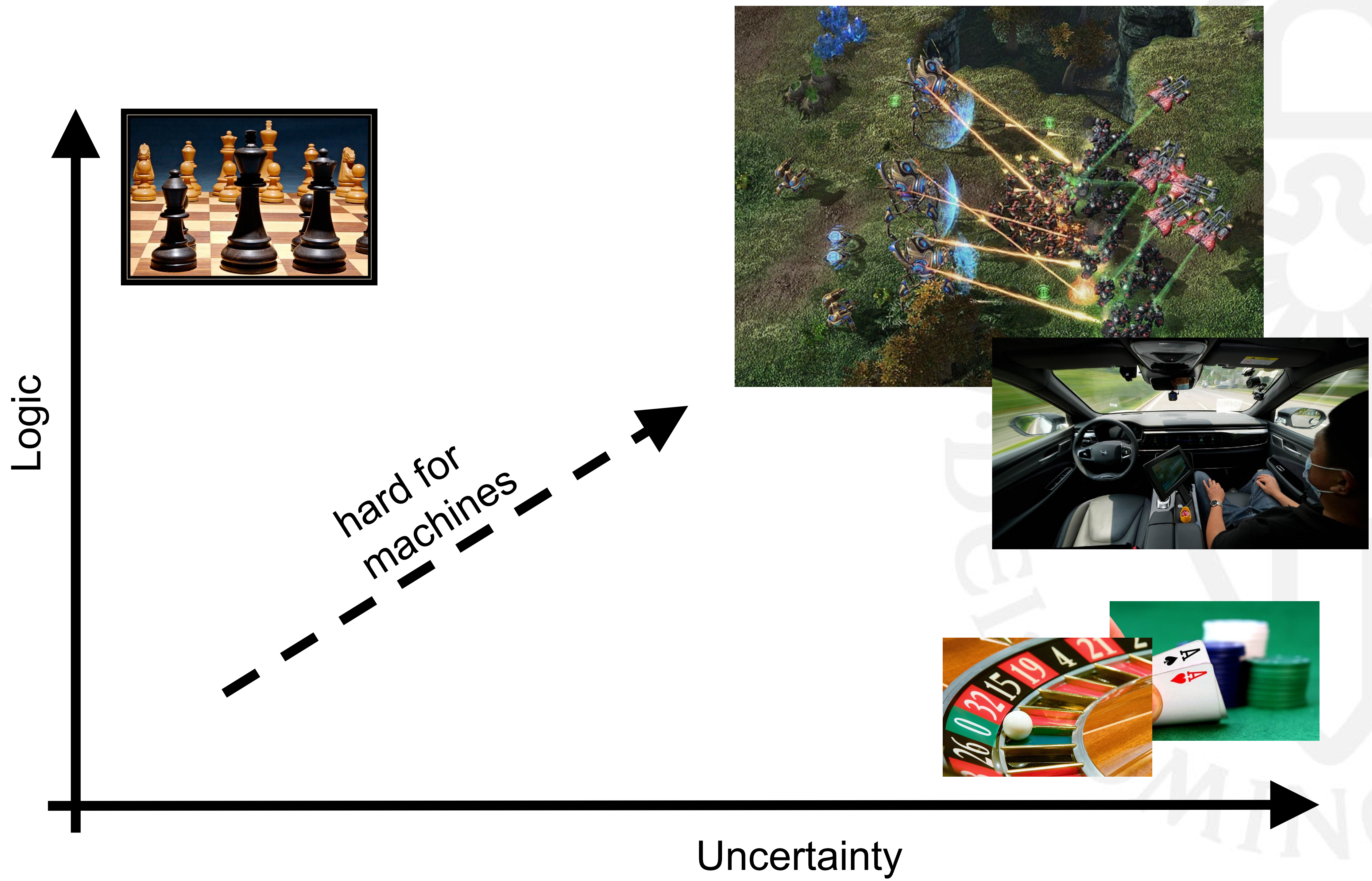
Marco Loog

Outline

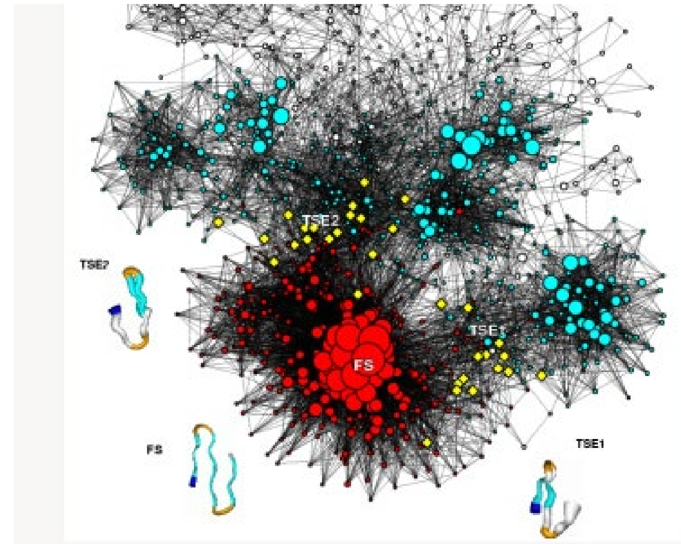
- Probability
- Statistics
- Hypothesis testing

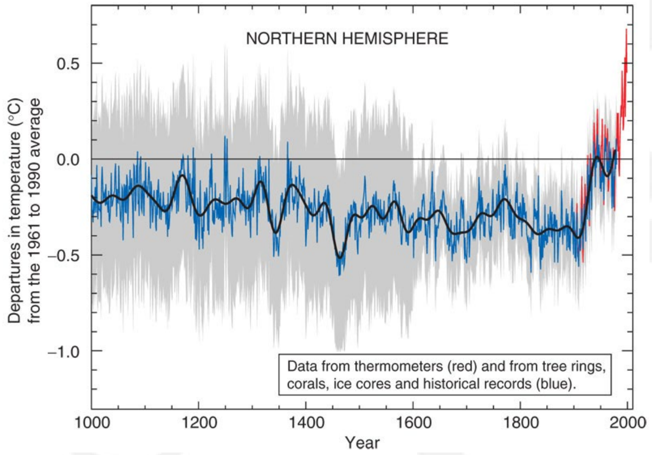
- Note : see Appendix C of TSK, first edition

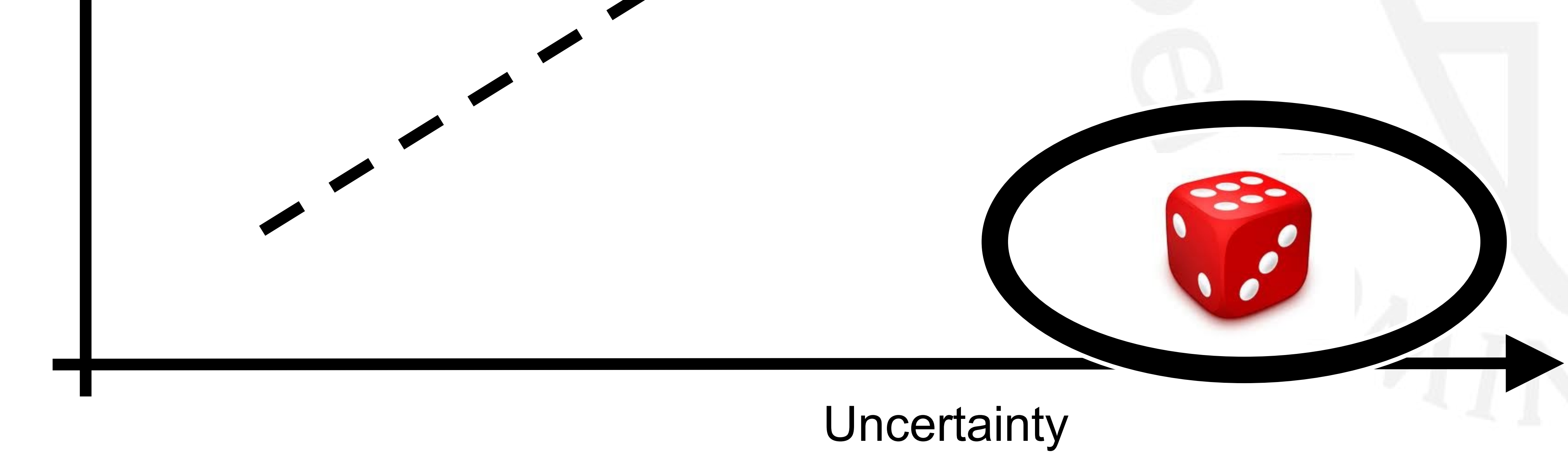
Probabilistic Reasoning



Probabilistic Reasoning





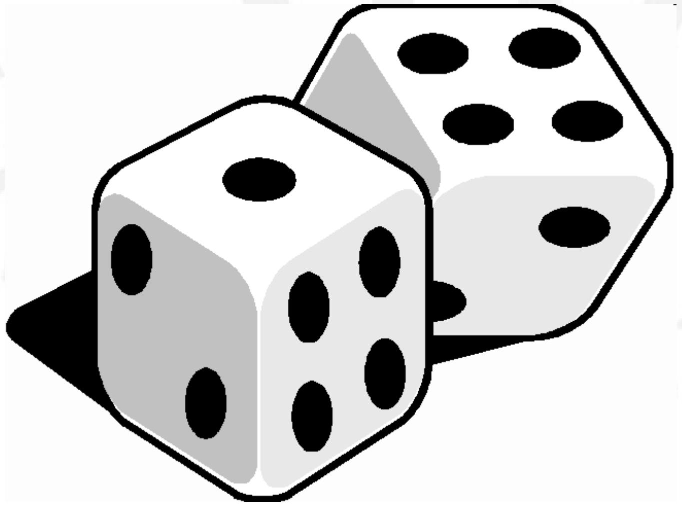


Concepts

- **Random experiment**
  - Rolling a die, flipping a coin, monitoring network traffic

- **Sample space**, all possible (single) outcomes : Ω
  - Ω = {1,2,3,4,5,6} for rolling a die
  - Ω = {heads,tails} for flipping a coin
  - $\Omega = [0,+\infty)$ 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 :

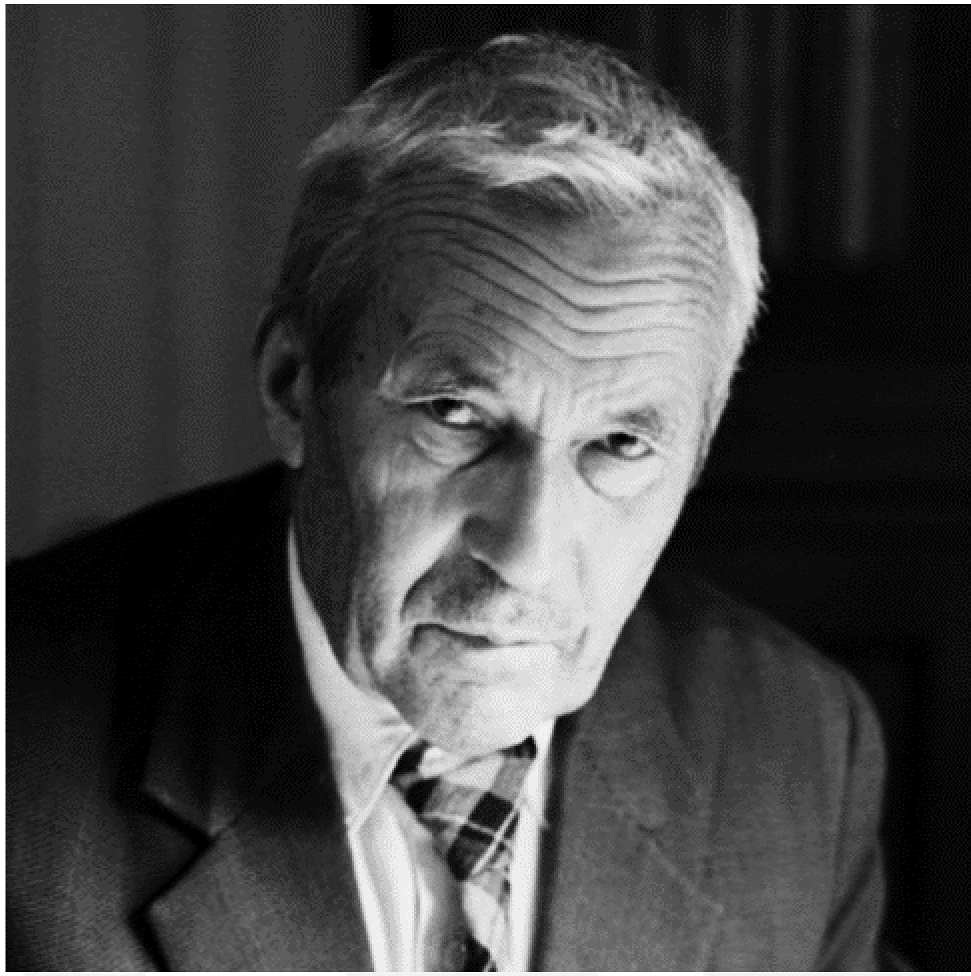
$$E \subseteq \Omega : 0 \leq P(E) \leq 1$$

  - 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
  - X equals number of heads when flipping a coin 30 times
  - X is time required to get back home

- Probability distribution [a.k.a. probability mass function] for a discrete random variable X :

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

Probability Distribution [Example]

- A fair die 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
$P(X)$					

Probability Density Function

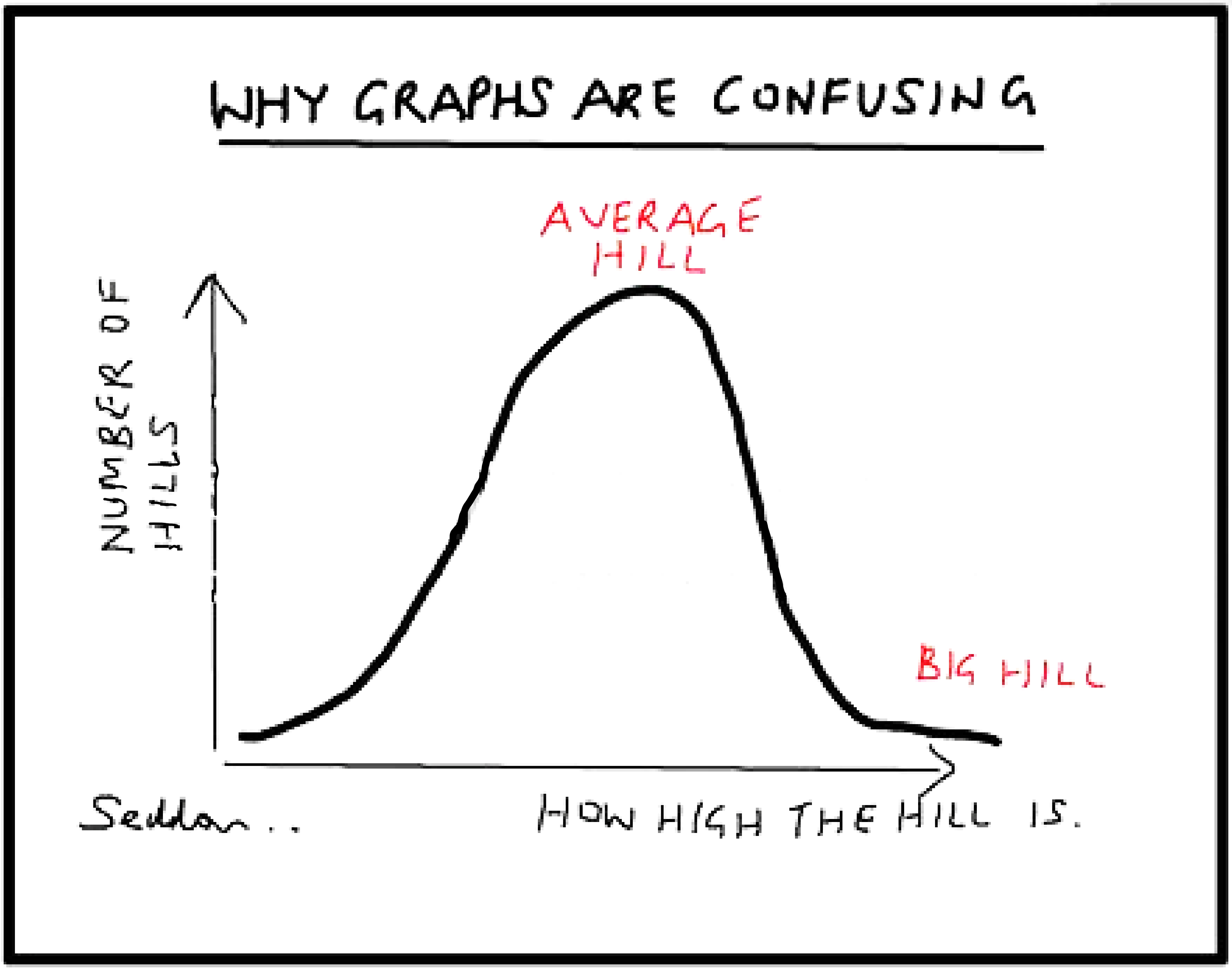
- For continuous variables :

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

- *f(x)* is called a probability density function

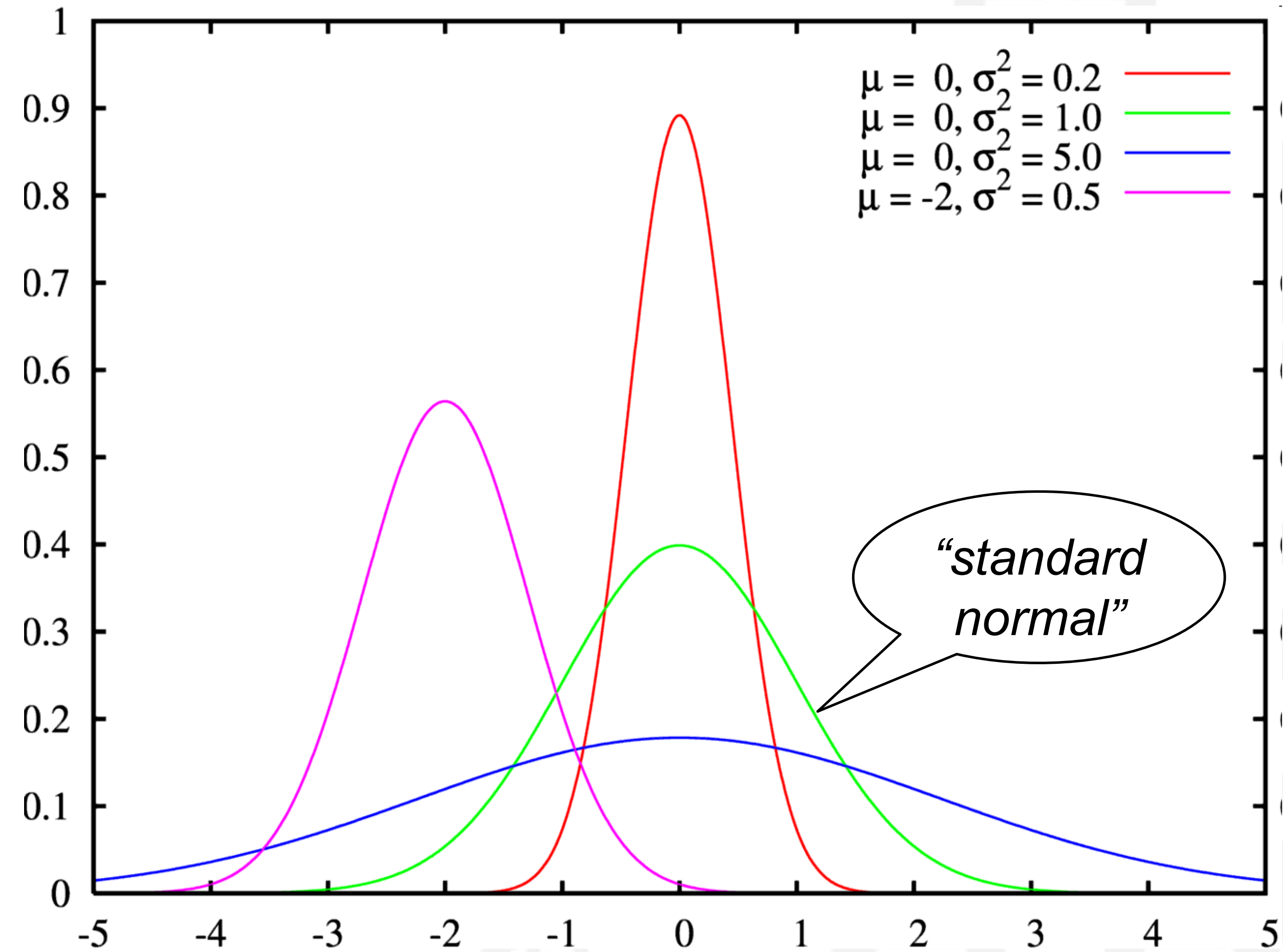
- Questions :

  - Probability that *X* takes a single, particular value equals?
  - Can *f(x)* be negative?
  - Can *f(x)* be larger than 1?


WHY GRAPHS ARE CONFUSING
AVERAGE HILL
NUMBER OF HILLS
BIG HILL
HOW HIGH THE HILL IS.

Gaussian Distribution

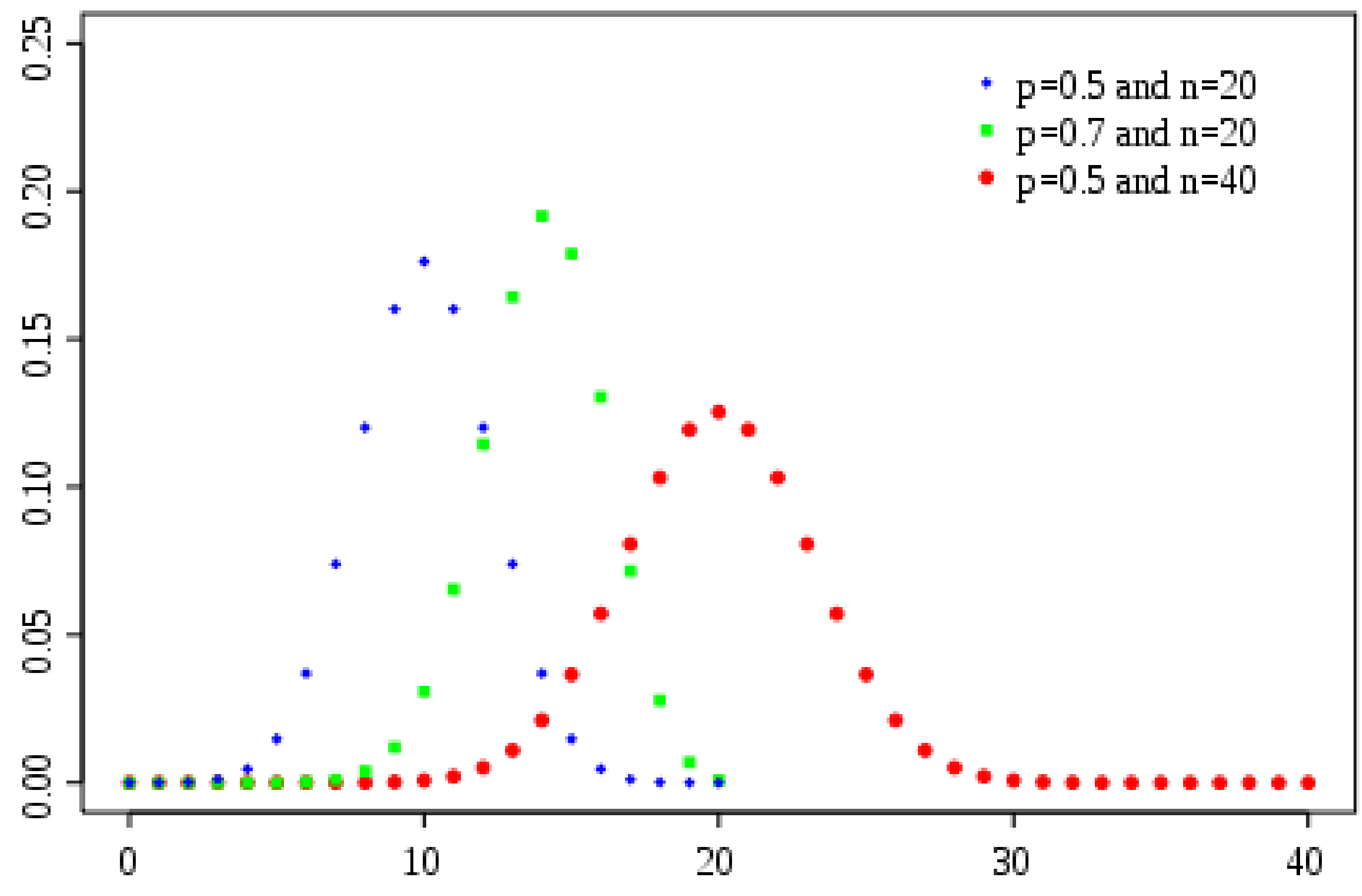
- Applicable in many fields due to **central limit theorem**
  - "Sum of many RVs is Gaussian"
  - Error / noise model
- Location parameter **[mean]** μ and spread **[standard deviation]** σ



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

Binomial Distribution

- Number of successes in a number of independent yes/no trials
  - Tossing a coin many times
  - # sixes in a game of dice



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

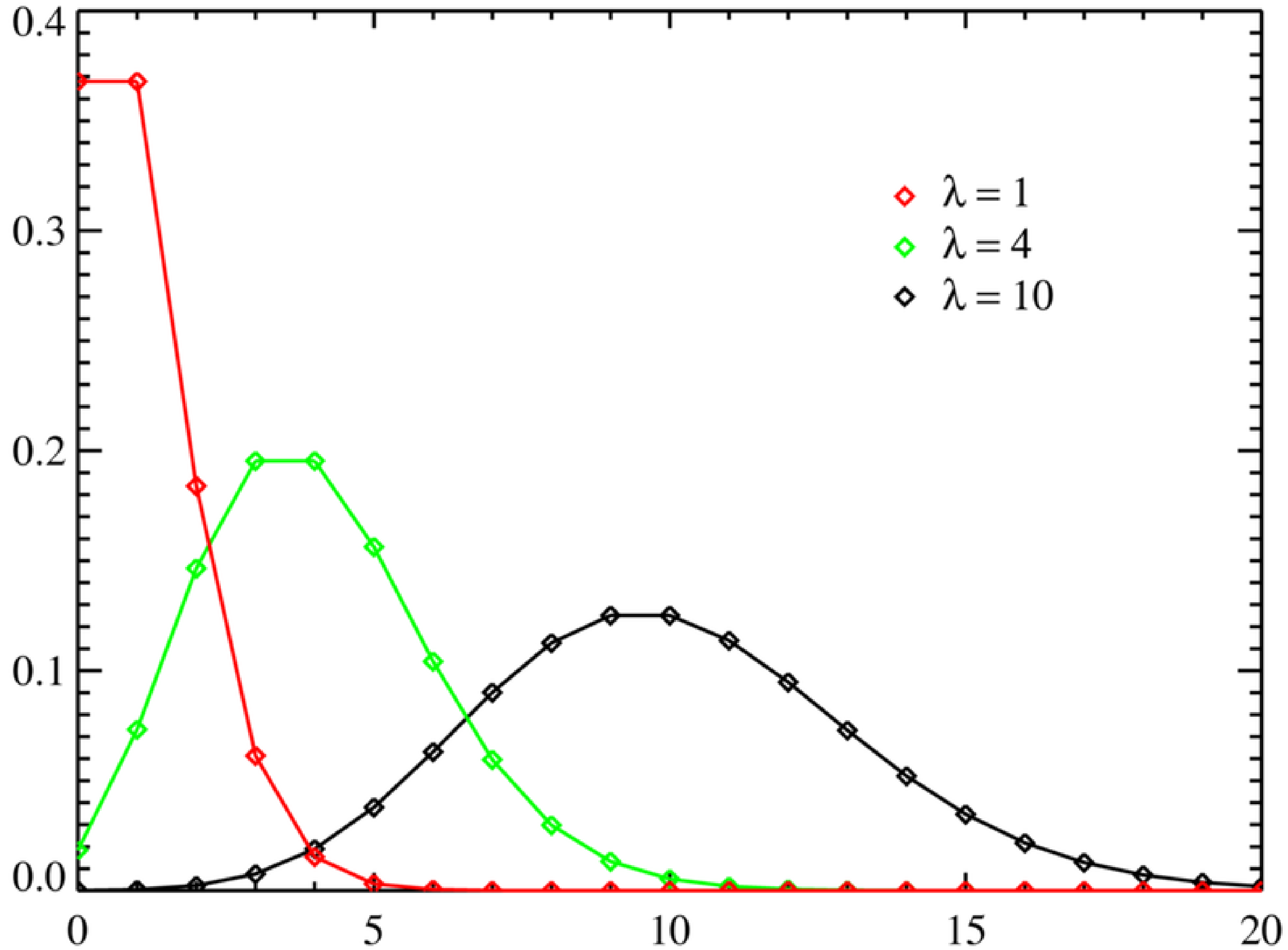
$$P(X=k;n,p)=\binom{n}{k}p^k\,(1-p)^{n-k}$$

Poisson Distribution

- Probability of number of events occurring in a fixed period of time/space
  - # people entering a building per hour
  - # hedgehogs killed per km of road
  - # mutations per 100.000 base pairs
  - # students dozing of per minute

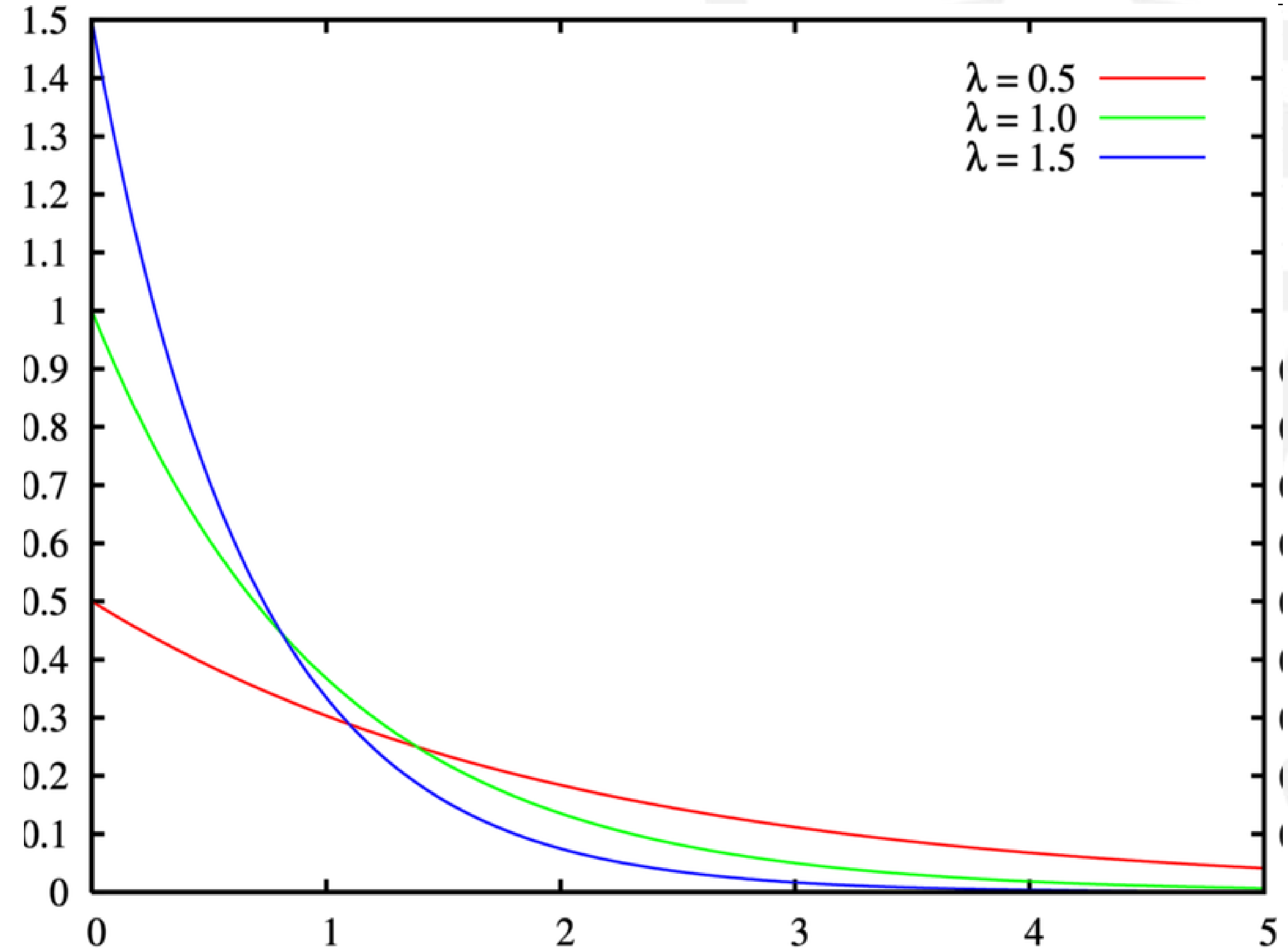
- **Rate** parameter λ

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



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

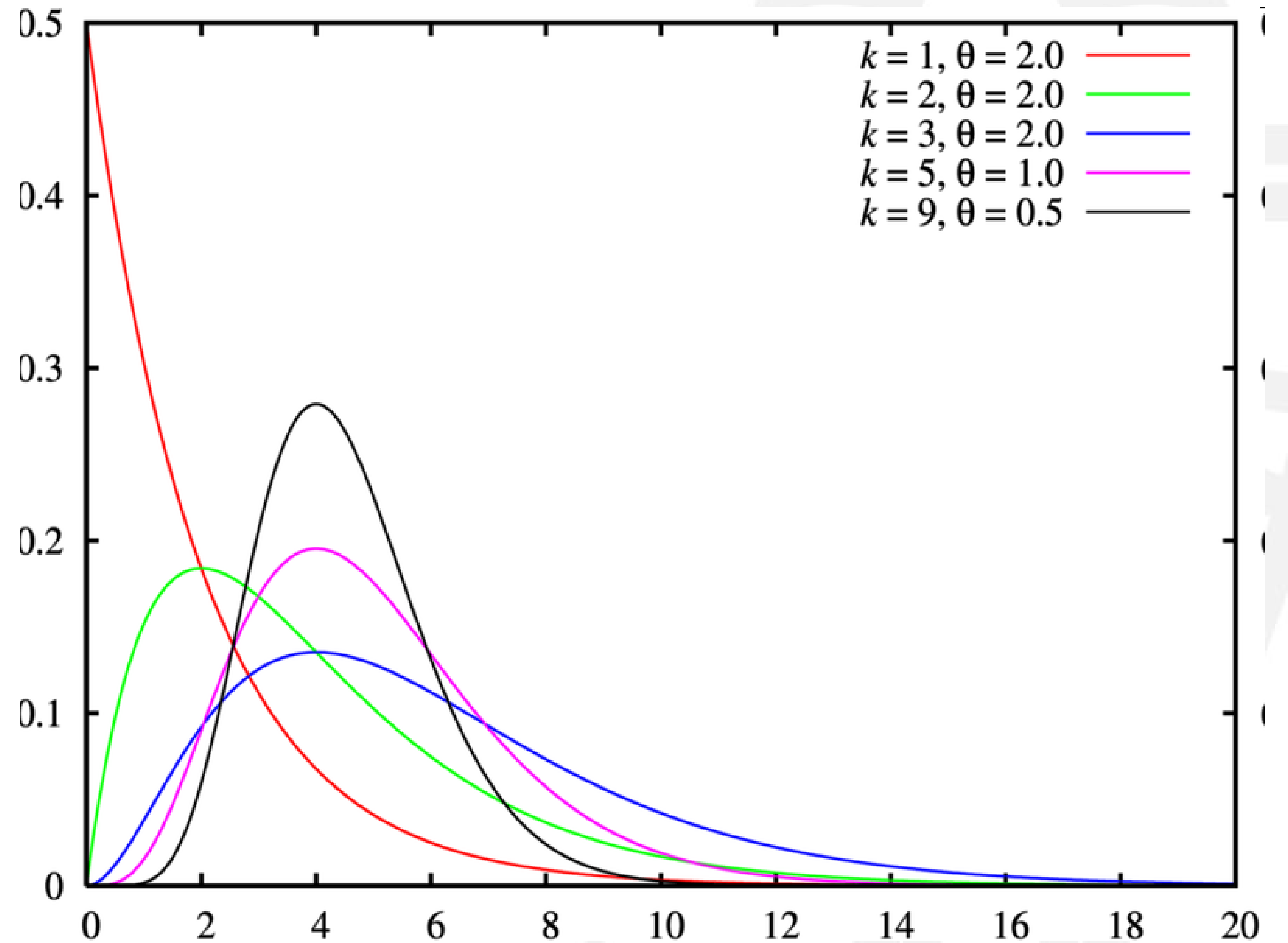


- **Rate** parameter λ

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

Gamma Distribution

- "Gaussian" for only positive values
  - Distribution of incomes
  - Lifetime of light bulbs



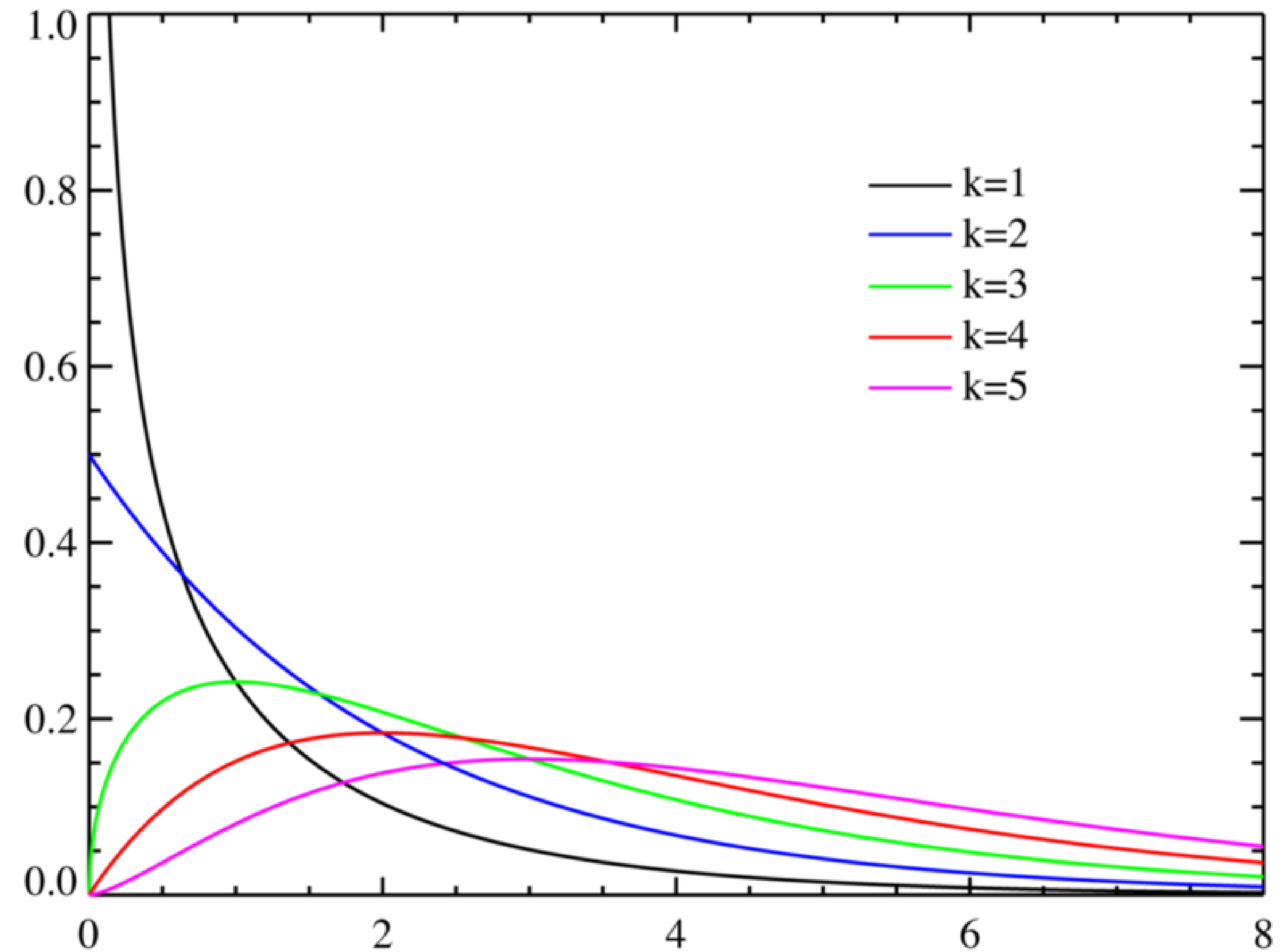
- **Scale** parameter θ and **shape** parameter k

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

Chi-square Distribution

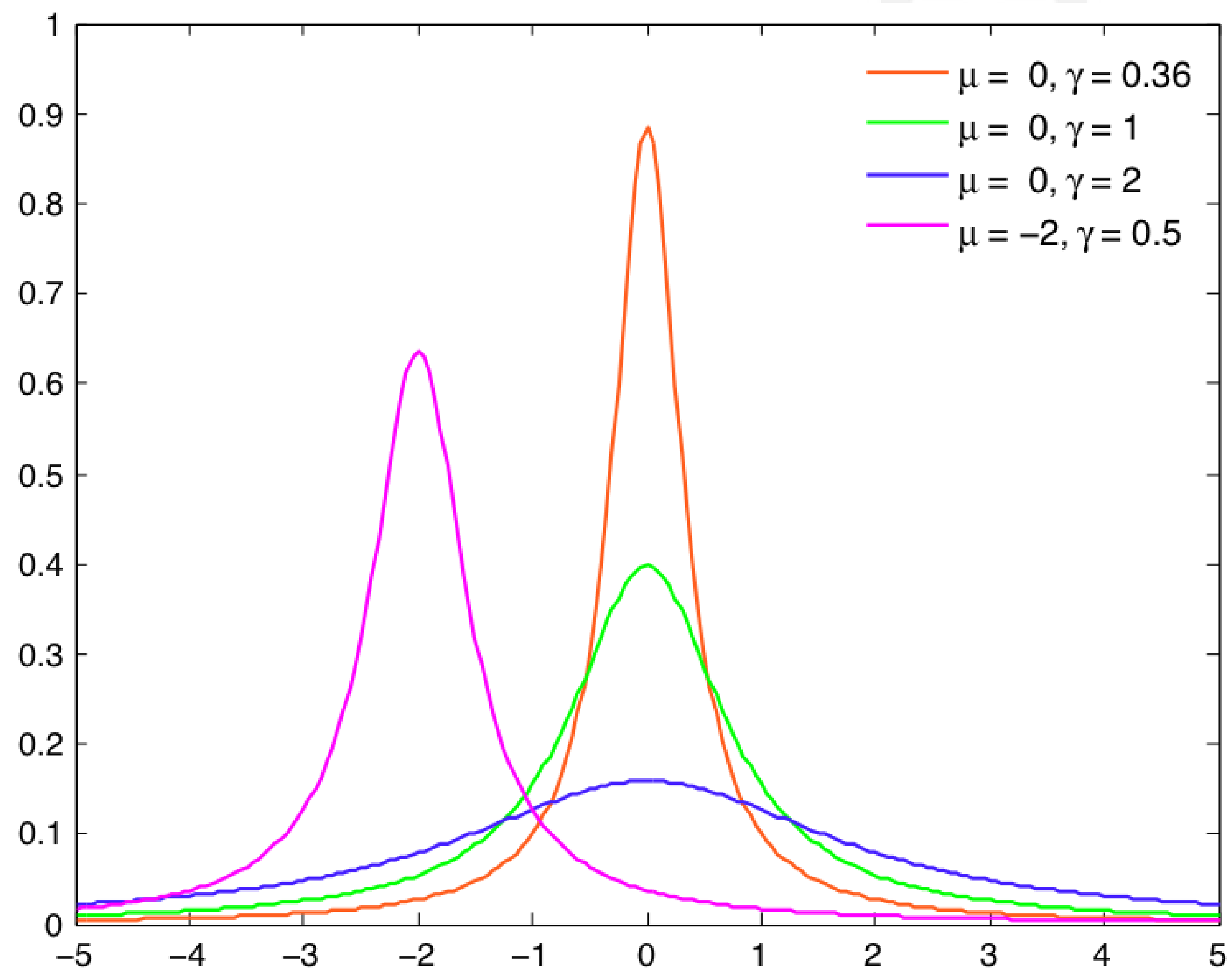
- Often used in statistical significance tests
- Special case of Gamma distribution [with $\theta \rightarrow 2$, $k \rightarrow k/2$]
- **Degrees of freedom** k : [distribution of sum of the squares of k normally distributed random variables]

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



Tails Matter…

- Cauchy distribution
  - Looks like a "fat tailed" Gaussian…
  - … but has no mean and no variance[!]
  - Very insensitive to outliers



- Location parameter μ and scale parameter γ

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

Stuff with Multiple Variables

- And some expectations...

Multiple Random Variables

- If X and Y are two RVs, 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 die
  - 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 die
  - *X* : outcome of die is 3 or higher
  - *Y* : even outcome

⇒ What is $P(Y|X)$?

- Direct : $P(Y|X) = P(\{4,6\} \mid \{3,4,5,6\}) = ½$
- Formula : $P(Y|X) = (P(X,Y) = ⅓) / (P(X) = ⅔) = ½$

Bayes' Theorem

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

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

- 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 die?

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

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 [i.e. uniform] 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 :

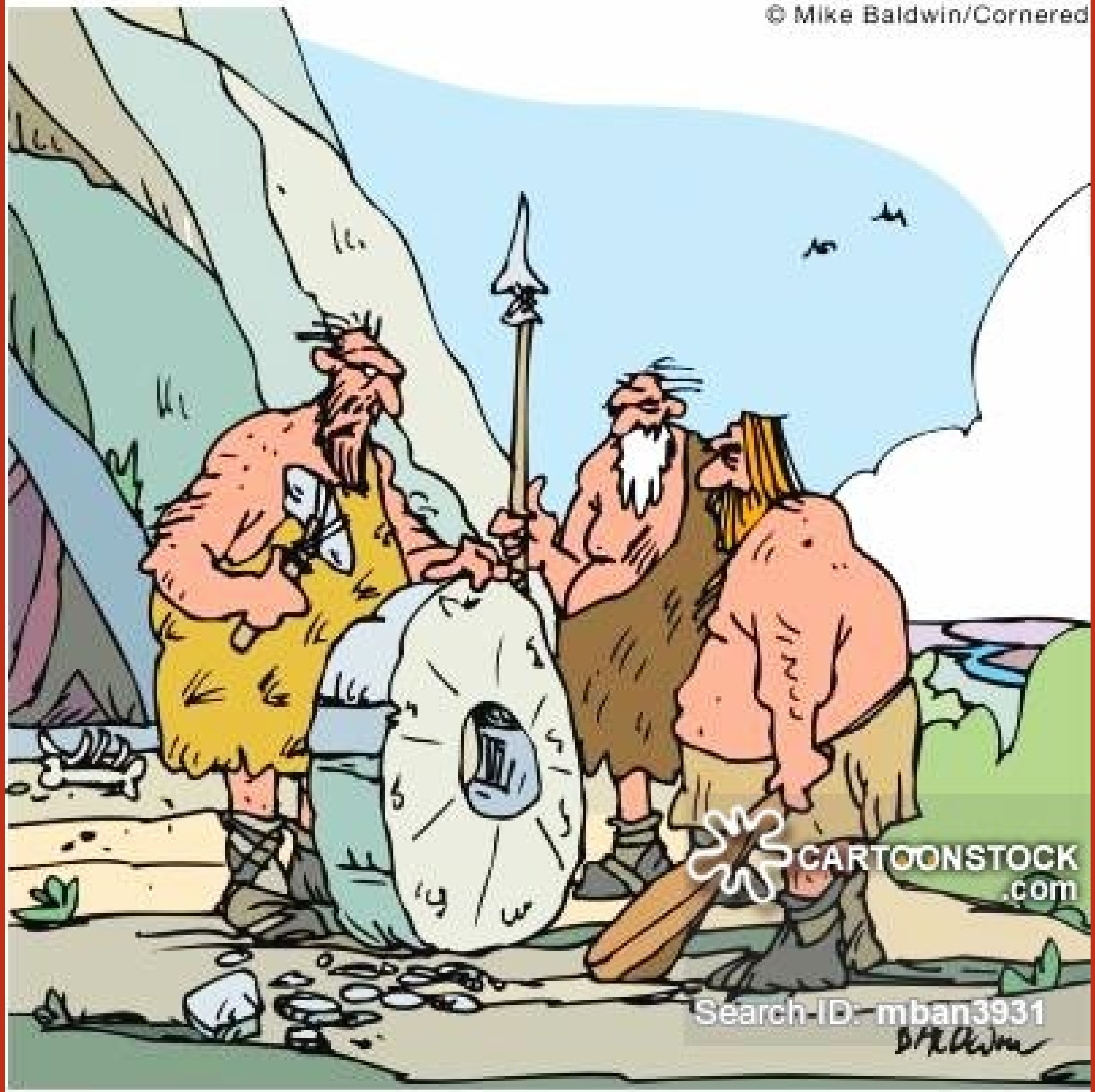
$$\mu_X = E[X] = \sum_k k\, P(X=k) \;\text{ or }\; \mu_X = \int x\, f(x)\, dx$$

- Variance :

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

- Covariance :

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



“If this works, it’ll change everything.
We could open a casino.”

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 is our best estimate for μ_X and σ_X^2?
- Obvious[?] choices :
  - Sample mean

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

  - Sample variance

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

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

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

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

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

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

Sample Variance

- From previous slide :

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

- From definitions and independent samples :

$$\sigma_X^2=E[X^2]-\mu_X^2$$
$$Cov[X,Y]=E[XY]-\mu_X\mu_Y$$

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

- 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

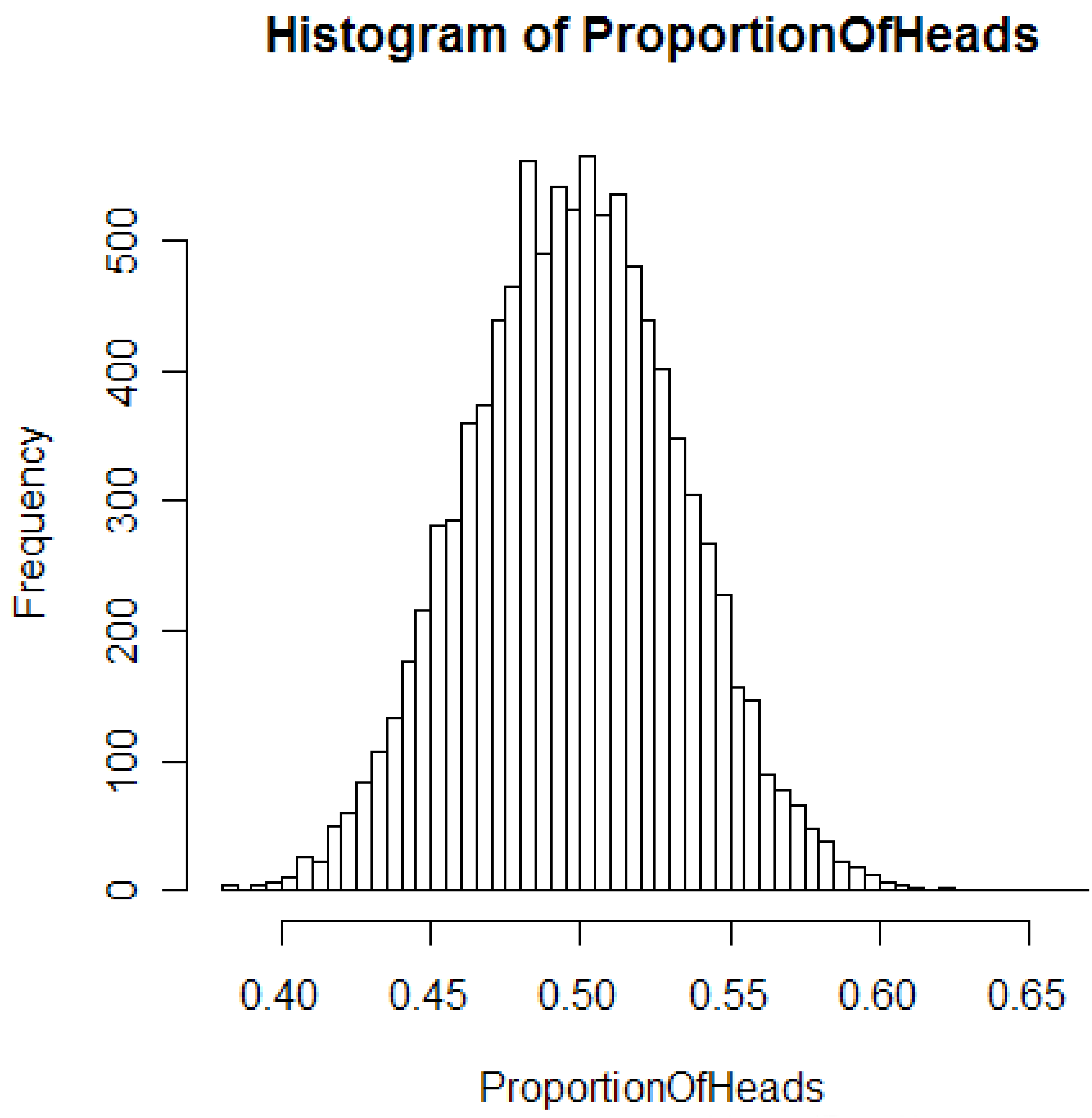
$$E\left[\left(\bar{X}-\mu_X\right)^2\right]=\frac{1}{N}\sigma_X^2$$

- 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 $\bar{X}$ of N samples from some distribution with **mean** μ_X and **variance** $\sigma_X{}^2$
- For large N, the distribution of the sample mean $\bar{X}$ approaches a **Gaussian** with mean μ_X and variance $\sigma_X{}^2/N$
- This is **independent** of the underlying distribution of the samples!



Confidence Intervals

- Would like to say a bit more than just our best [point] 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 $\alpha = 0.1$ 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

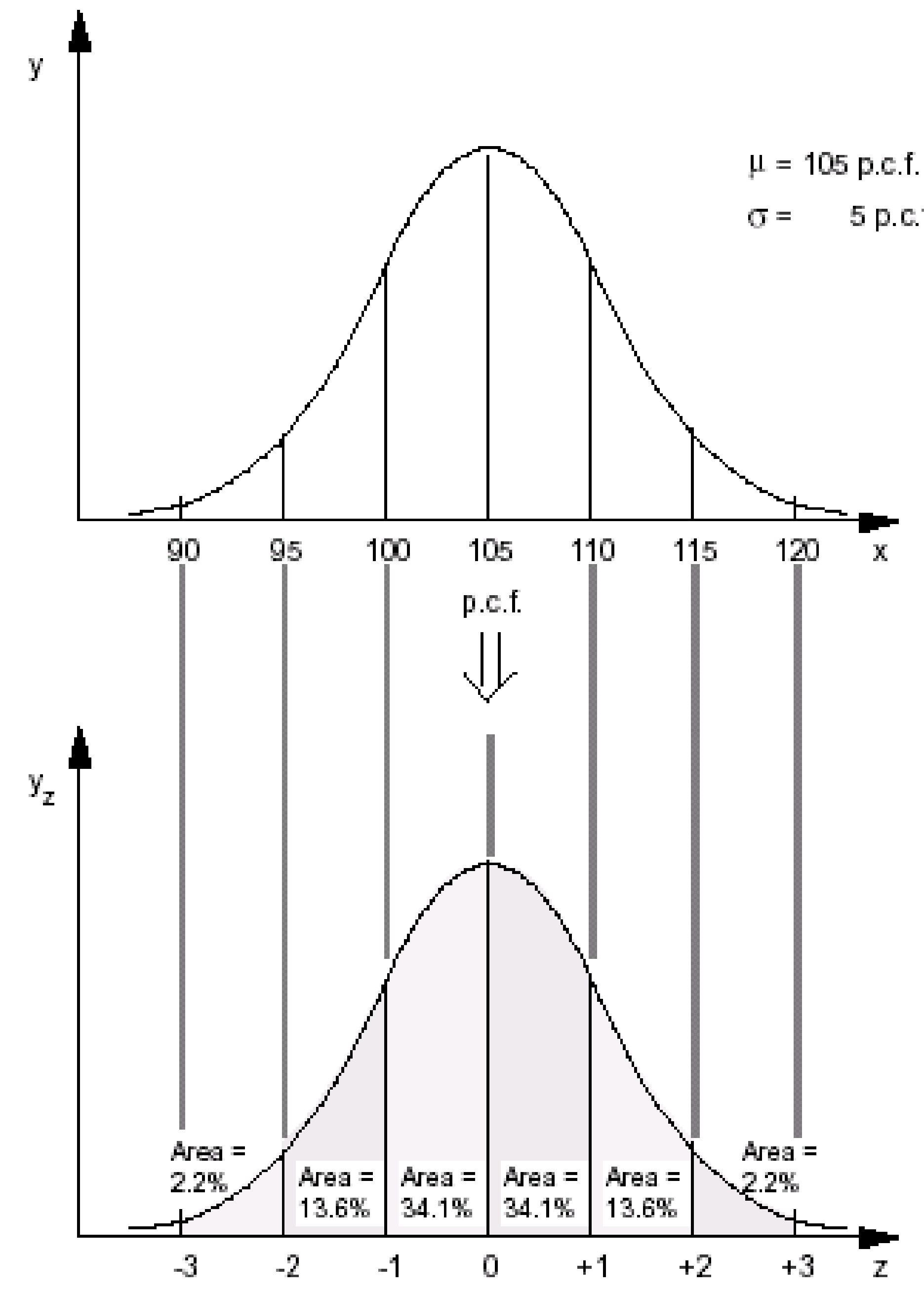
- **Central limit theorem** : the distribution of the population mean $\overline{X}$ approaches a normal distribution with mean μ_X and variance $\sigma_X{}^2/N$

- That is, the variable

$$Z = \frac{\overline{X} - \mu_X}{\sigma_X / \sqrt{N}}$$

has a **standard normal** distribution [mean 0, variance 1] :

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



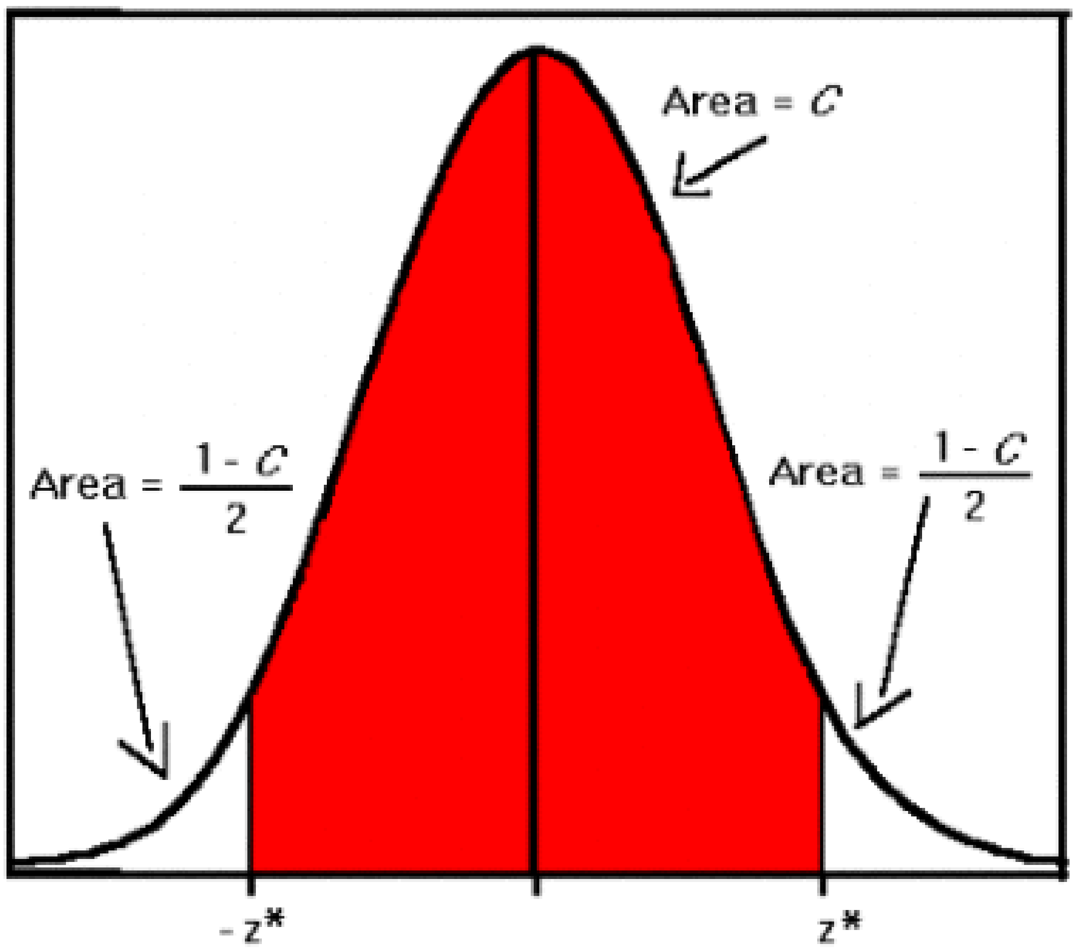
Confidence Interval for Sample Mean

- Rewriting : **observe** a sample mean $\bar{x}$, the confidence interval for μ_X reads

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

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

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



Hypothesis Testing

- Should we **accept** or **reject** a hypothesis [e.g., “Barbie is more clever than Ken”] 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 $\alpha = 0.05$
  - Following this “rule” blindly is not necessarily a good idea…

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?

- Example null hypotheses :
  - Coin / die is fair
  - No difference between classification methods
  - Random variables X and Y are independent

- Example test statistics :
  - 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

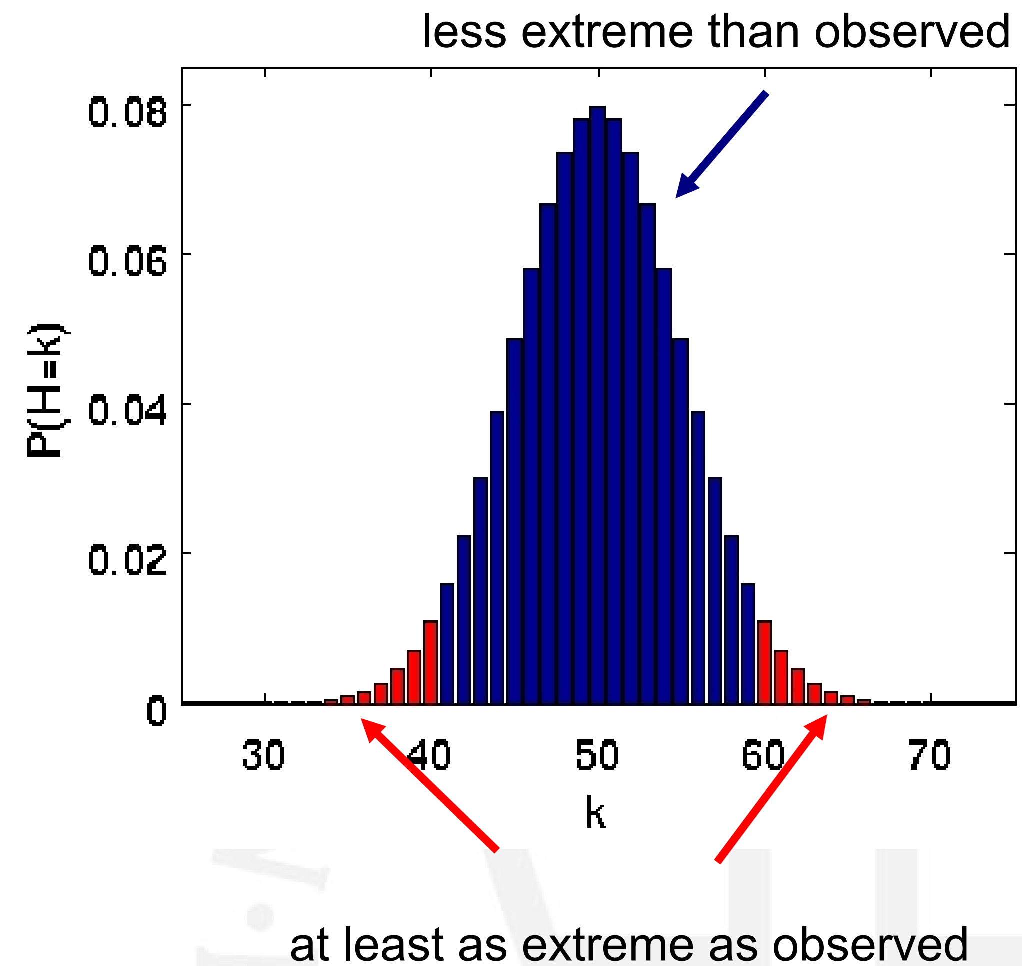
- Formulate the *null* ["simple"] *hypothesis*
- Define a *significance level α*
- Define a *test statistic θ* with a known probability distribution under the null hypothesis
- Compute the value of *θ* from the observed data
- Compute the **confidence interval** for *θ* under the null hypothesis for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

Example : Fair Coin

- 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?

"Do we want to just flip a coin, or hire expensive consultants so they can flip a coin?"

Example : Fair Coin



- 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
  - That is, no [or not enough] reason to reject the null hypothesis

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]

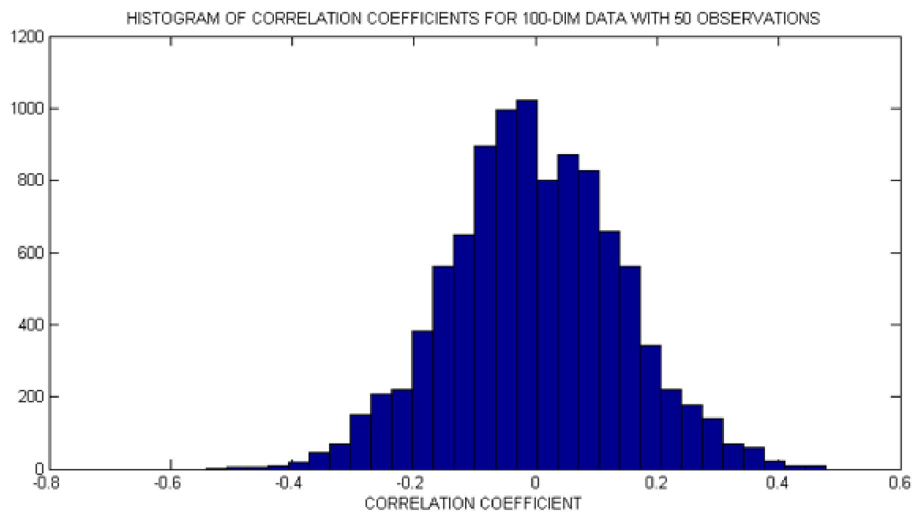
Publication Bias and p-Value Hunting/p-Hacking

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

Think back of First Lecture…

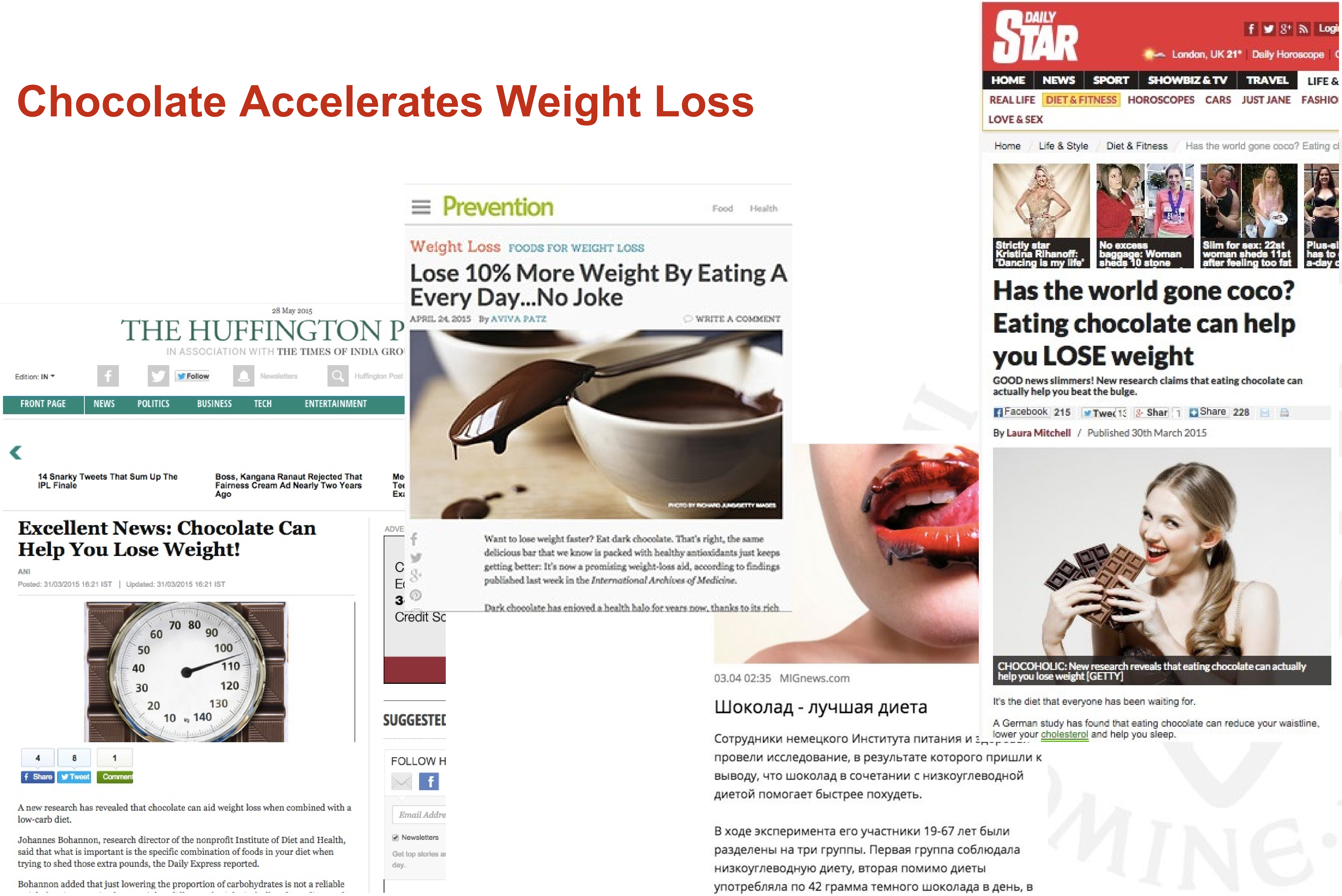
Rough Explanation of "Data Fishing"

- Data set with
  - 50 data vectors
  - 100 variables
  - Even if data are entirely random [no dependence] there is a very high probability some variables will appear dependent just by chance.



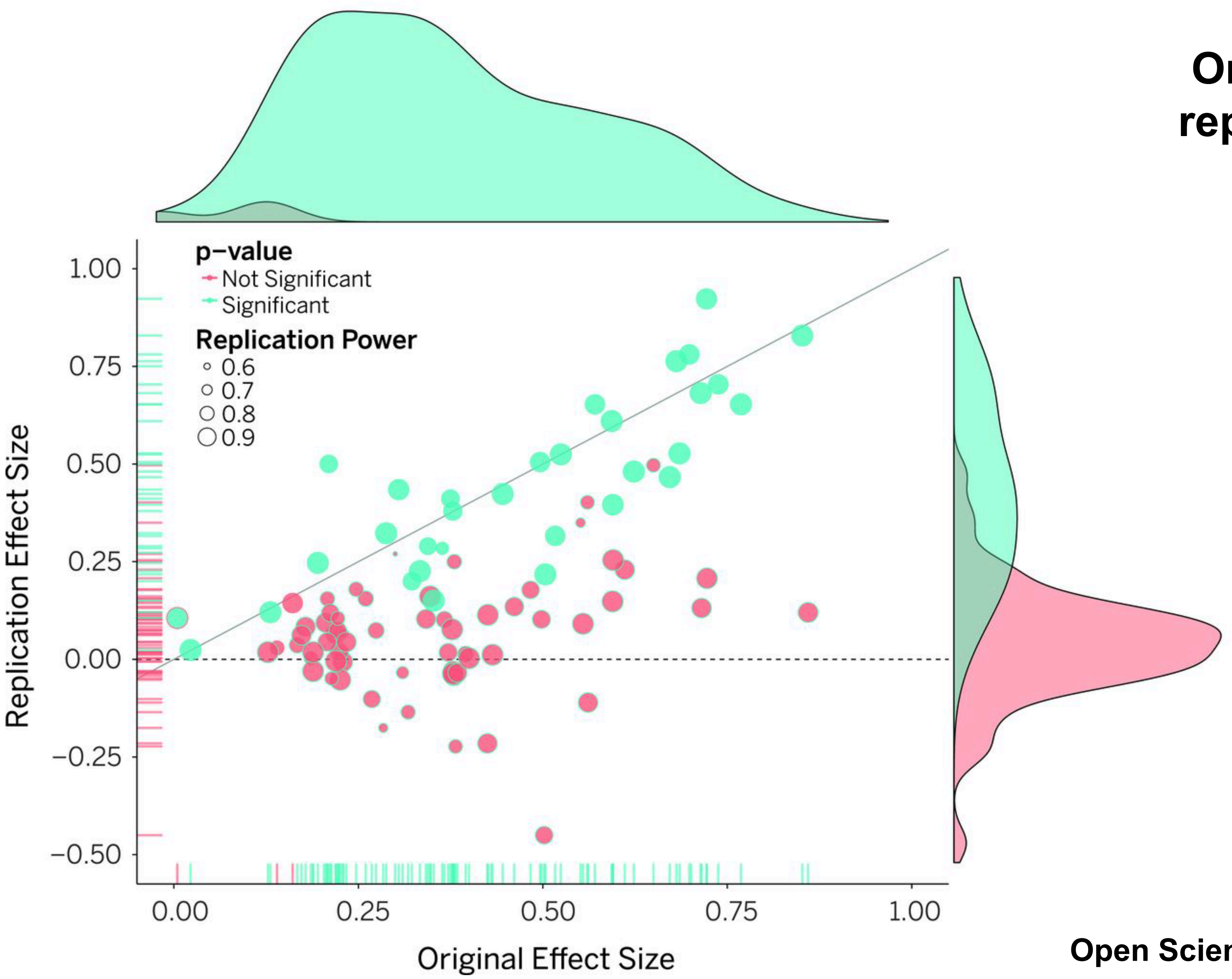
Radboud University Nijmegen

Chocolate Accelerates Weight Loss



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

Reproducibility of Psychological Science

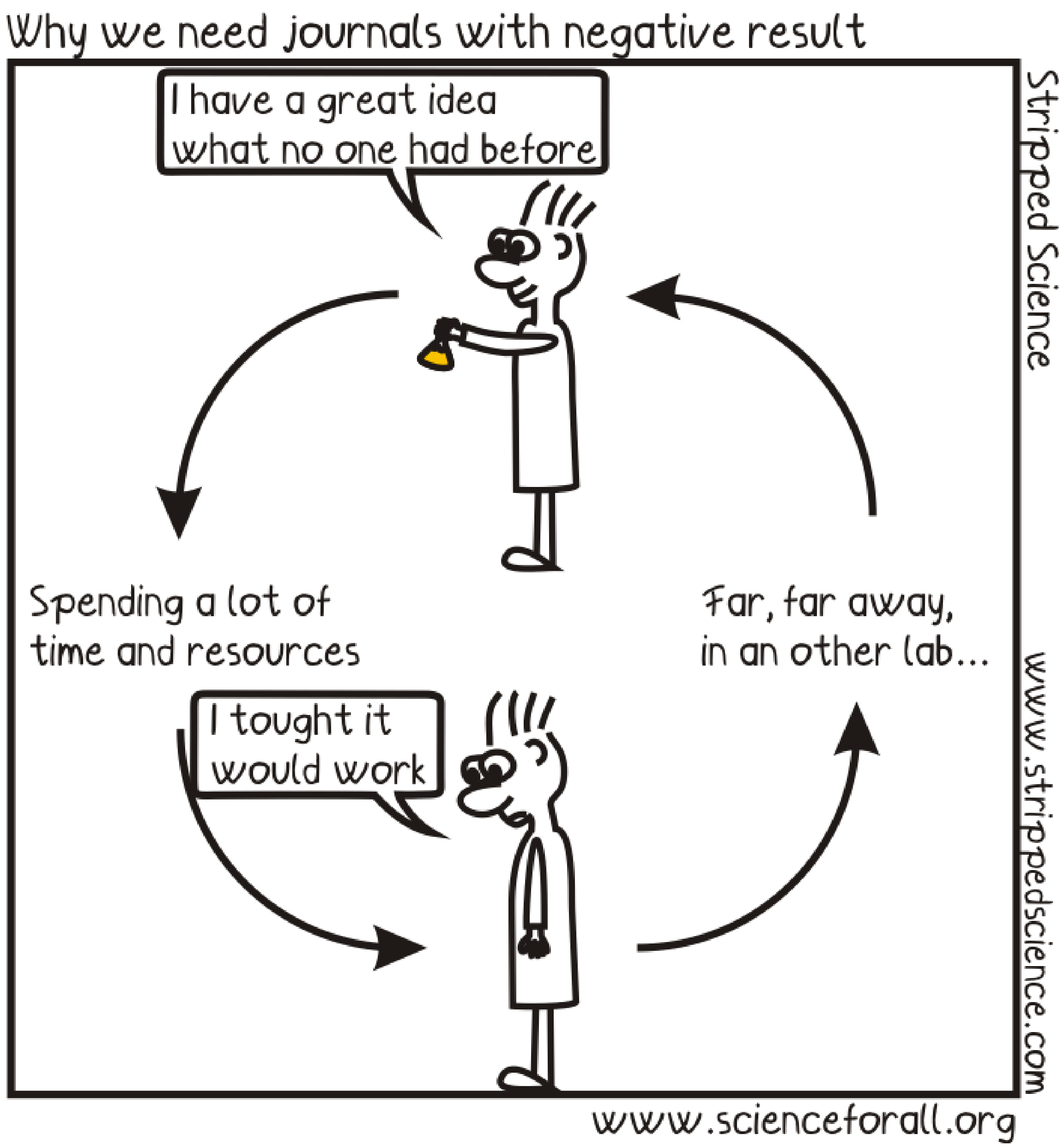


Original study effect size versus replication effect size (correlation coefficients).

Open Science Collaboration Science 2015;349:aac4716

Published by AAAS

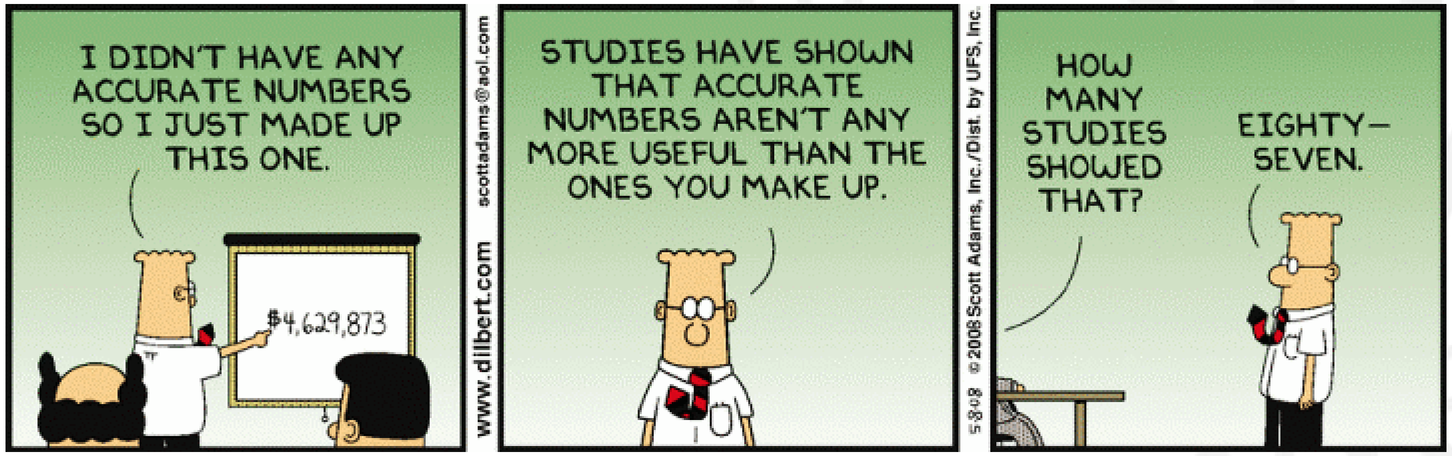
Publishing Negative Results



The Null Journal
Negative Results, Failures, and Voids

JOURNAL OF NEGATIVE RESULTS IN BIOMEDICINE

Qs?


I DIDN'T HAVE ANY ACCURATE NUMBERS SO I JUST MADE UP THIS ONE.
$4,629,873
scottadams@aol.com
www.dilbert.com
STUDIES HAVE SHOWN THAT ACCURATE NUMBERS AREN'T ANY MORE USEFUL THAN THE ONES YOU MAKE UP.
© 2008 Scott Adams, Inc./Dist. by UFS, Inc.
HOW MANY STUDIES SHOWED THAT?
EIGHTY-SEVEN.