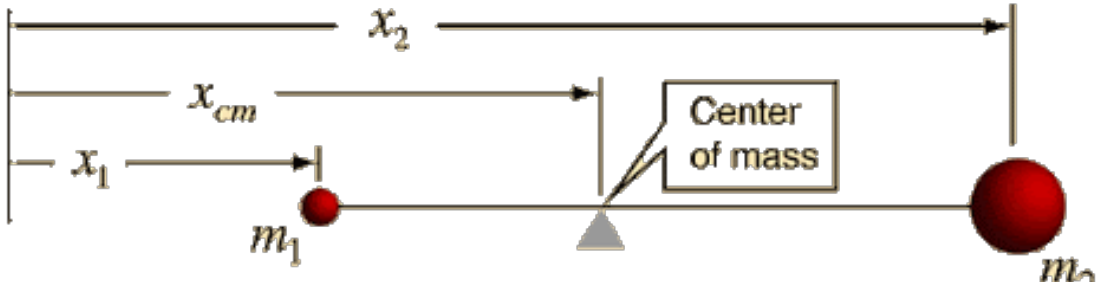
# CS 215
# Data Analysis and Interpretation

## Expectation

Suyash P. Awate

# Expectation

- "Expectation" of the random variable; "Expected value" of the random variable; "**Mean**" of the random variable.
- "Expected value" isn't the value that is most likely to be observed in the random experiment
- Indicates the center of mass of the probability mass/density function




For two masses: $x_{cm} = \dfrac{m_1 x_1 + m_2 x_2}{m_1 + m_2}$

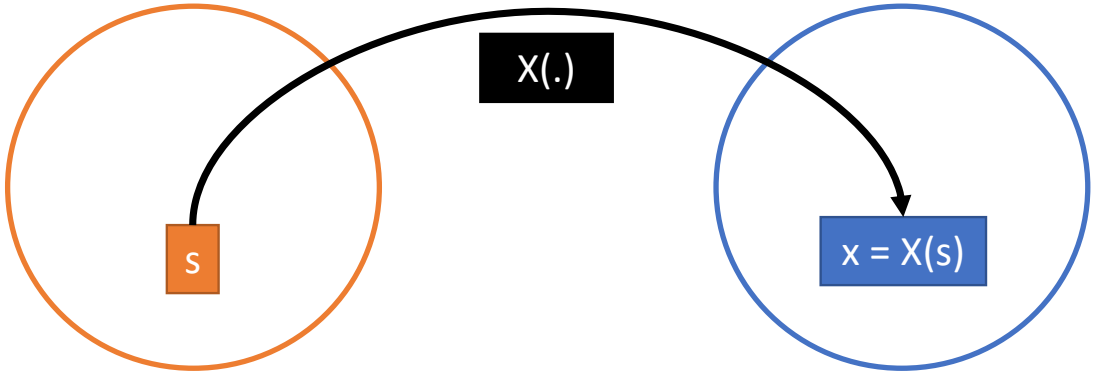
# Expectation

- Definition:
  **Expectation** of a **Discrete** Random Variable: $E[X] := \sum_i x_i P(X = x_i)$
- Frequentist interpretation of probabilities and expectation
  - If a random experiment is repeated infinitely many times,
    then the proportion of time that event E occurs is P(E)
  - If a random experiment underlying a discrete random variable X
    is repeated infinitely many times,
    then the fraction of the number of instances when X takes value x is P(X = x)
  - So, in $N \to \infty$ experiments, $N.P(X=x_i)$ number of times we will get $X=x_i$
  - So, arithmetic average of all observed values $x_i$
    across all $N \to \infty$ experiments is
    $(1/N) \sum_i (x_i) (N.P(X=x_i))$
    $= E[X]$

# Expectation

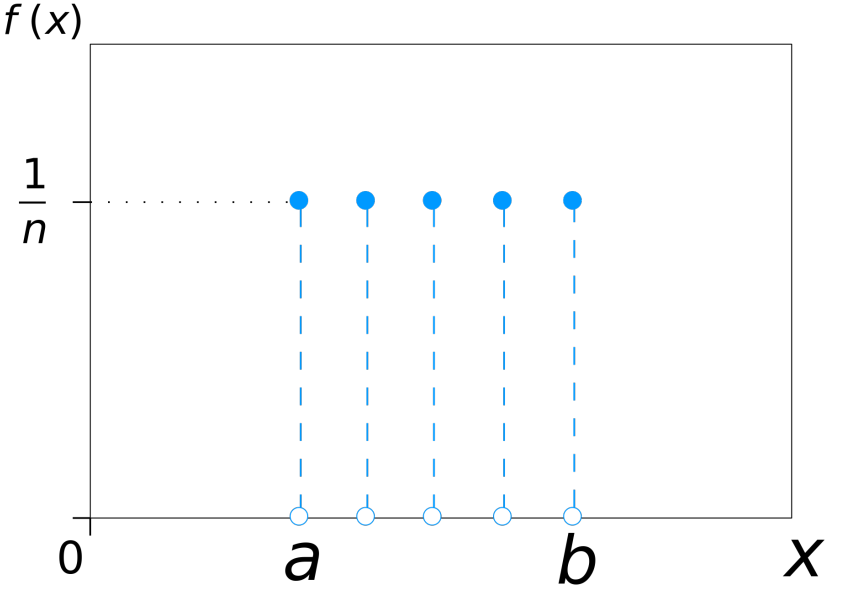
- Another Formulation of Expectation
  - Recall:
    - Random variable X is a function defined on a probability space {Ω,ß,P}
    - Function X:Ω→R, maps each element in sample space Ω to a single numerical value belonging to the set of real numbers

$$E[X] := \sum_i x_i P(X = x_i)$$
$$= \sum_i x_i \left( \sum_{s \in \Omega : X(s) = x_i} P(s) \right)$$
$$= \sum_{s \in \Omega} X(s) P(s)$$



# Expectation

- Example
  - “Expected value” for the uniform random variable modelling die roll
    - Values on die are {1,2,3,4,5,6}
    - E[X] = 3.5
  - **Expectation of a uniform random variable (discrete case)**
    - If X has uniform distribution over n consecutive integers over [a,b], then E[X] = (a+b)/2

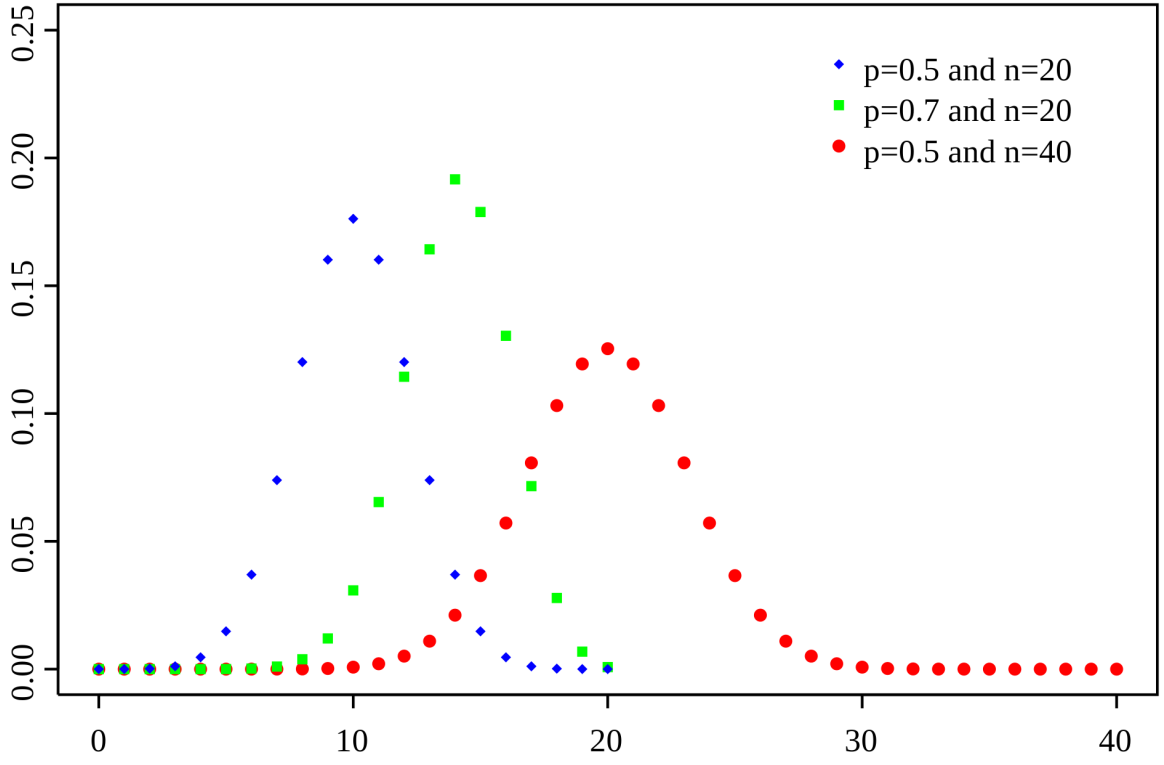


# Expectation

- Example
  - **Expectation of a binomial random variable** (when n=1, this is **Bernoulli**)

$$
\begin{aligned}
\mathrm{E}(X) &= \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} \\
&= \sum_{k=1}^{n} k \binom{n}{k} p^k q^{n-k} \\
&= \sum_{k=1}^{n} n \binom{n-1}{k-1} p^k q^{n-k} \\
&= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} q^{(n-1)-(k-1)} \\
&= np \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \\
&= np
\end{aligned}
$$

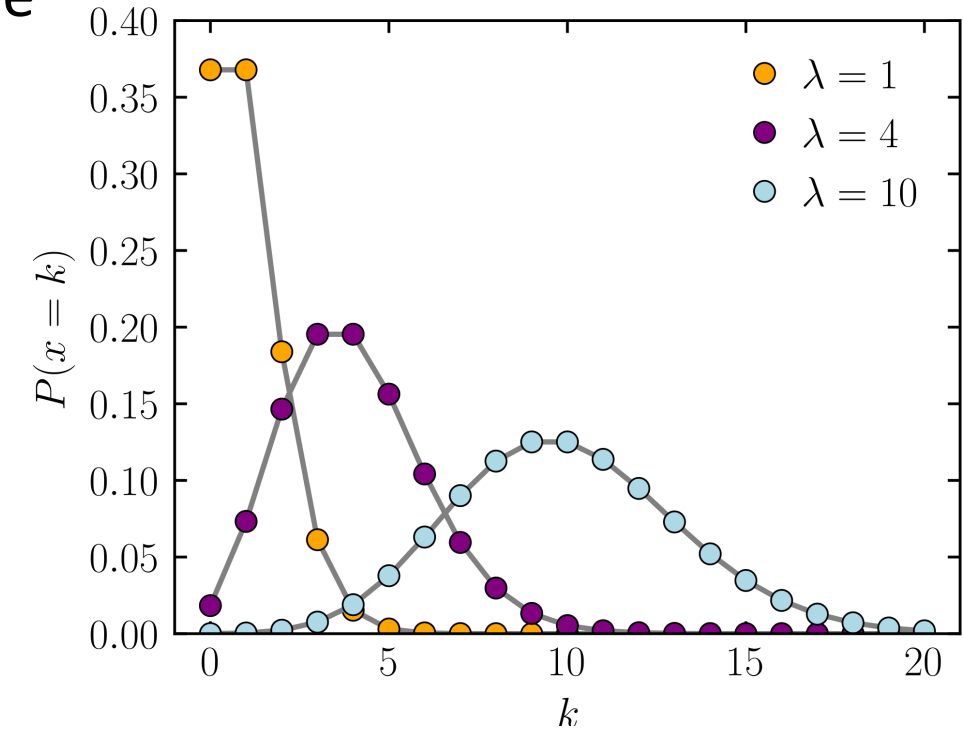
j := k – 1
m := n – 1



# Expectation

- Example
  - **Expectation of a Poisson random variable**
  - Consider random events/arrivals occurring at a constant average rate λ>0, i.e., λ events/arrivals (typically) per unit time

$$
\begin{aligned}
\mathsf{E}\,(X) &= \sum_{k \geq 0} k \frac{1}{k!} \lambda^k e^{-\lambda} \\
&= \lambda e^{-\lambda} \sum_{k \geq 1} \frac{1}{(k-1)!} \lambda^{k-1} \\
&= \lambda e^{-\lambda} \sum_{j \geq 0} \frac{\lambda^j}{j!} \\
&= \lambda e^{-\lambda} e^{\lambda} \\
&= \lambda
\end{aligned}
$$



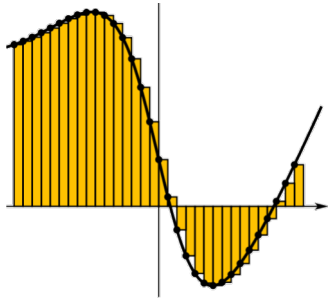

  - This gives meaning to parameter λ as average number of events in unit time

# Expectation

- Definition:

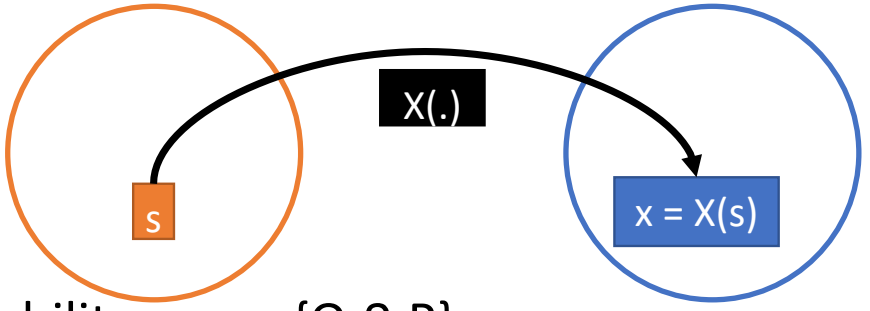
  **Expectation** of a **Continuous** Random variable: $E[X] := \int_{-\infty}^{\infty} xP(x)dx$

- Frequentist interpretation of probabilities and expectation

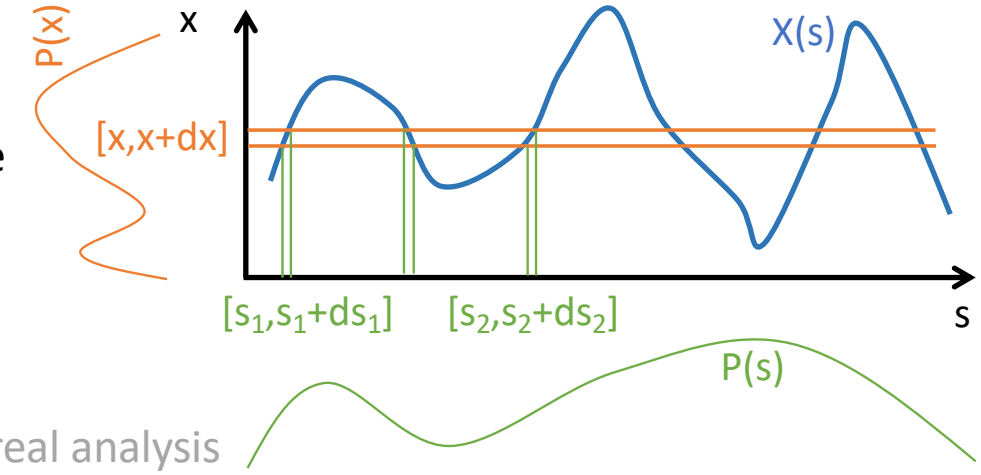


  - If a random experiment underlying a continuous random variable X is repeated $N \to \infty$ times,
    then,
    for a tiny interval [x,x+dx],
    the proportion of time X takes values within interval is approx. N.(dx.P(x))
  - So, in $N \to \infty$ experiments,
    approximately N.(dx.P($x_i$)) number of times we will get X within [$x_i$,$x_i$+dx]
  - So, arithmetic average of all observed X values across all intervals [$x_i$,$x_i$+dx] across all $N \to \infty$ experiments is $(1/N) \sum_i (x_i) (N.dx.P(x_i))$
  - In the limit that $dx \to 0$, this average→E[X]

# Expectation




- Another Formulation of Expectation
  - Recall:
    - Random variable X is a function defined on a probability space {Ω,ß,P}
    - Function X:Ω→R, maps each element in sample space Ω to a single numerical value belonging to the set of real numbers
    - $E[X] := \int_{-\infty}^{\infty} xP(x)dx = \int_{-\infty}^{\infty} X(s)P(s)ds$
      - Intuition remains the same as in the discrete case
      - Using probability-mass conservation: P(x)dx is approximated by $P(s_1)ds_1 + P(s_2)ds_2 + \ldots$
      - Thus, x.P(x)dx is approximated by $X(s_1).P(s_1)ds_1 + X(s_2).P(s_2)ds_2 + \ldots$
      - A more rigorous proof needs advanced results in real analysis

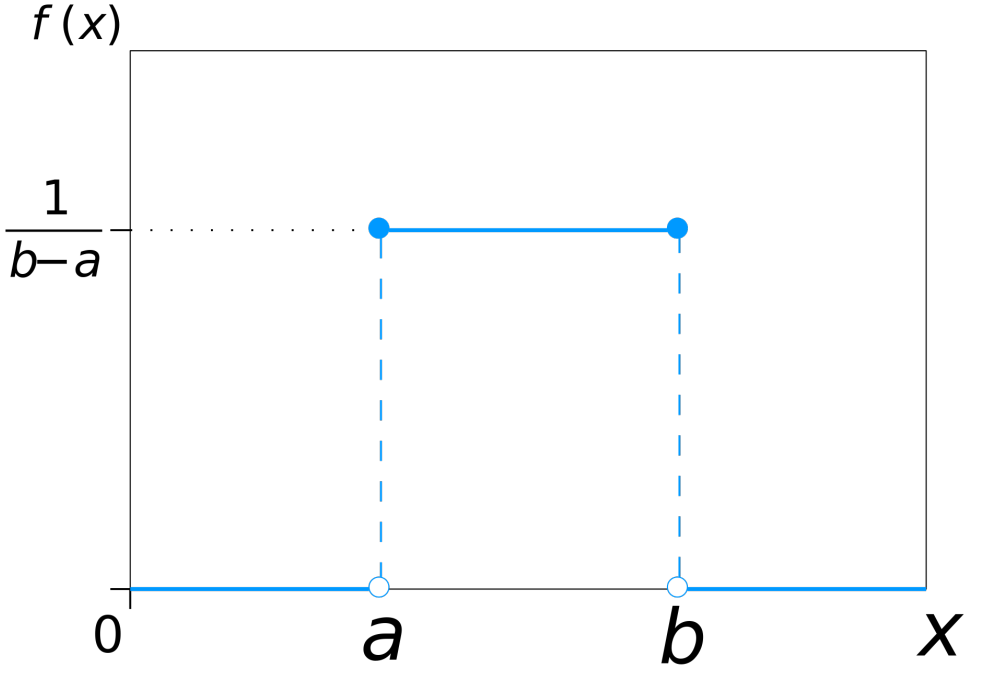



$$E[X] := \sum_i x_i p(X = x_i) = \sum_i x_i \left( \sum_{s \in \Omega : X(s) = x_i} P(s) \right) = \sum_{s \in \Omega} X(s)P(s)$$

# Expectation

- Example
  - **Expectation of a uniform random variable (continuous case)**

$$\begin{aligned} \mathrm{E}(X) &= \int_{-\infty}^{\infty} x f(x)\, dx = \int_a^b x \frac{1}{b-a}\, dx = \frac{1}{2(b-a)} \Big[ x^2 \Big]_a^b \\ &= \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{b+a}{2} \end{aligned}$$



# Expectation

**PDF**

P(x) = 0, for all x < 0

$P(x) = \lambda \exp(-\lambda x), \forall x \geq 0$

**CDF**

f(x) = 0, for all x < 0

$f(x) = 1 - \exp(-\lambda x), \forall x \geq 0$

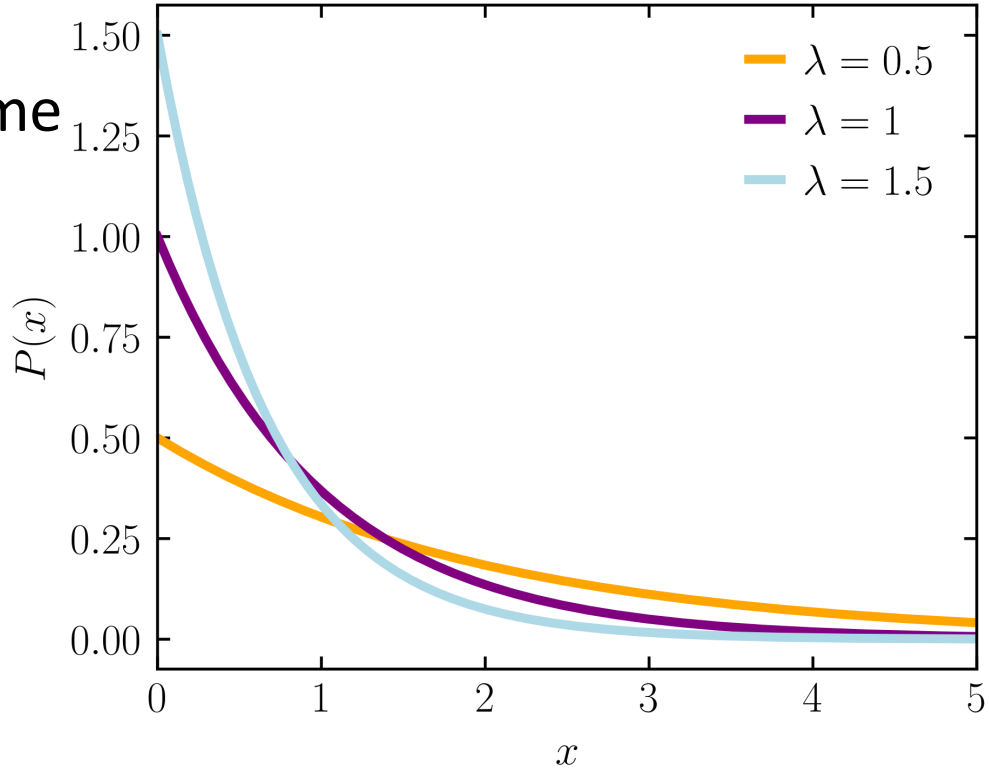
- Example
  - **Expectation of an exponential random variable**
  - Consider random events/arrivals occurring at a constant average rate λ > 0, i.e., λ events/arrivals (typically) per unit time
  - Define **β := 1/λ**

$$\mathrm{E}(X) = \int_0^\infty x \frac{1}{\beta} \exp\left(-\frac{x}{\beta}\right) \mathrm{d}x$$

$$= \beta \int_0^\infty u \exp(-u)\, \mathrm{d}u \qquad u = \frac{x}{\beta}$$

$$= \left[-\beta (u+1) \exp(-u)\right]_0^\infty$$

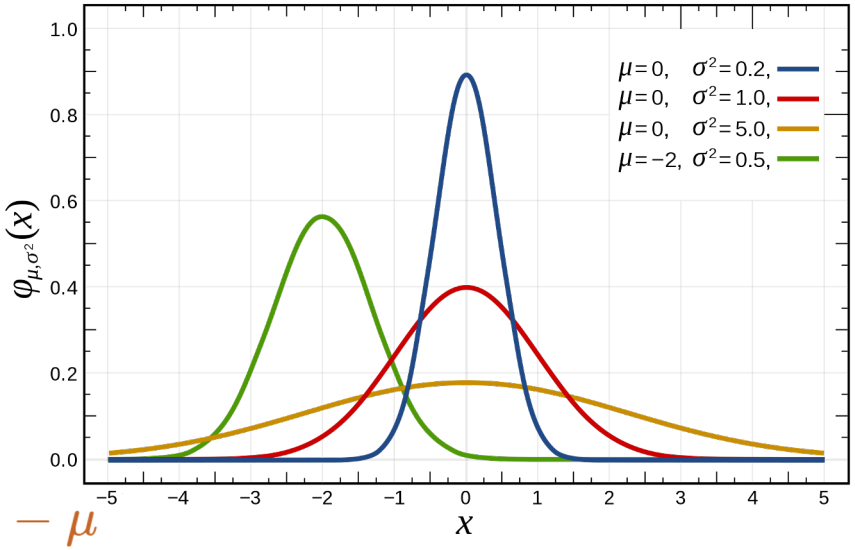
$$= \beta$$




- This gives meaning to parameter β as expected/average inter-arrival time

# Expectation

- Example
  - **Expectation of a Gaussian random variable**




$$
\begin{aligned}
\mathrm{E}\,(X) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x \\
&= \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sqrt{2}\sigma t + \mu) \exp\left(-t^2\right) \mathrm{d}t \qquad t = \frac{x-\mu}{\sqrt{2}\sigma} \\
&= \frac{1}{\sqrt{\pi}} \left(\sqrt{2}\sigma \int_{-\infty}^{\infty} t \exp\left(-t^2\right) \mathrm{d}t + \mu \int_{-\infty}^{\infty} \exp\left(-t^2\right) \mathrm{d}t\right) \\
&= \frac{1}{\sqrt{\pi}} \left(\sqrt{2}\sigma \left[-\frac{1}{2}\exp\left(-t^2\right)\right]_{-\infty}^{\infty} + \mu\sqrt{\pi}\right) \\
&= \frac{\mu\sqrt{\pi}}{\sqrt{\pi}} \\
&= \mu
\end{aligned}
$$

# Expectation

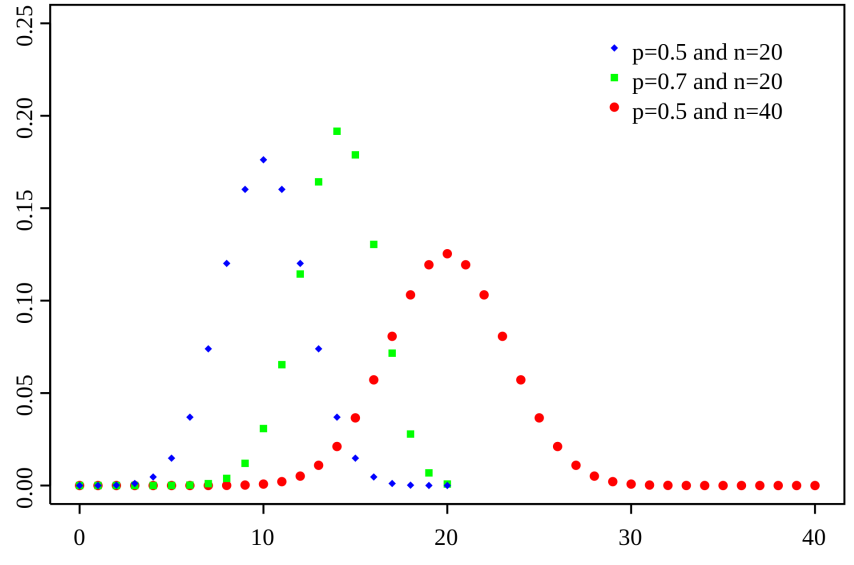
- Example
  - **Expectation of a limiting case of binomial**
  - As n tends to infinity,

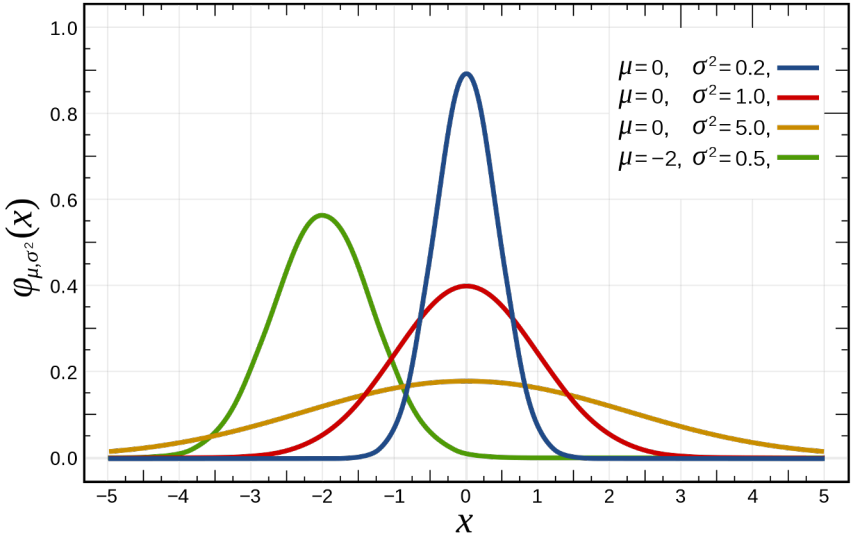
    binomial $f(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$

    tends to a

    "Gaussian" form $\frac{1}{\sqrt{2\pi npq}} e^{-(x-np)^2/2npq}$

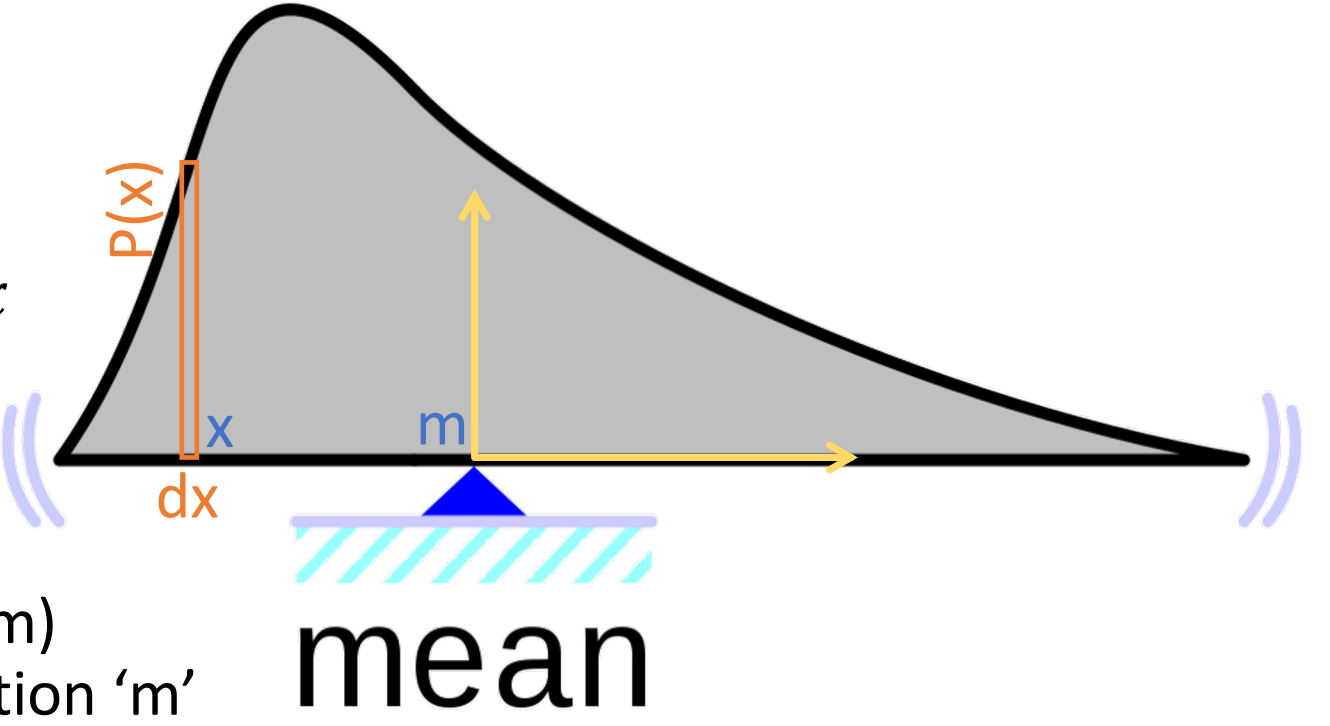
  - Gaussian expectation μ(=np here) is consistent with binomial expectation np



# Expectation

- **Mean as the center of mass**
  - By definition,
    mean m := E[X] := $\int_x xP(x)dx$
  - Thus, $\int_x (x-m)P(x)dx = 0$
  - Mass P(x)dx
    placed around location ‘x’
    applies a torque ∝ P(x)dx.(x-m)
    at the fulcrum placed at location ‘m’
    - Because the integral $\int_x (x-m)P(x)dx$ is zero,
      the net torque around the fulcrum ‘m’ is zero
    - Hence, ‘m’ is the center of mass



# Expectation

- **Linearity of Expectation**
  - For both discrete and continuous random variables
- For random variables X and Y sharing a probability space (Ω,ß,P), the following rules hold:
  - E[X + Y] = E[X] + E[Y]

$$E[X+Y] = \sum_{s\in\Omega}(X(s)+Y(s))P(s) = E[X]+E[Y]$$

  - E[X + c] = E[X] + c, where ‘c’ is a constant

$$E[X+c] = \sum_{s\in\Omega}(X(s)+c)P(s) = \sum_{s\in\Omega}X(s)P(s) + c\sum_{s\in\Omega}P(s) = E[X]+c$$

  - E[a X] = a E[X], where ‘a’ is a scalar constant

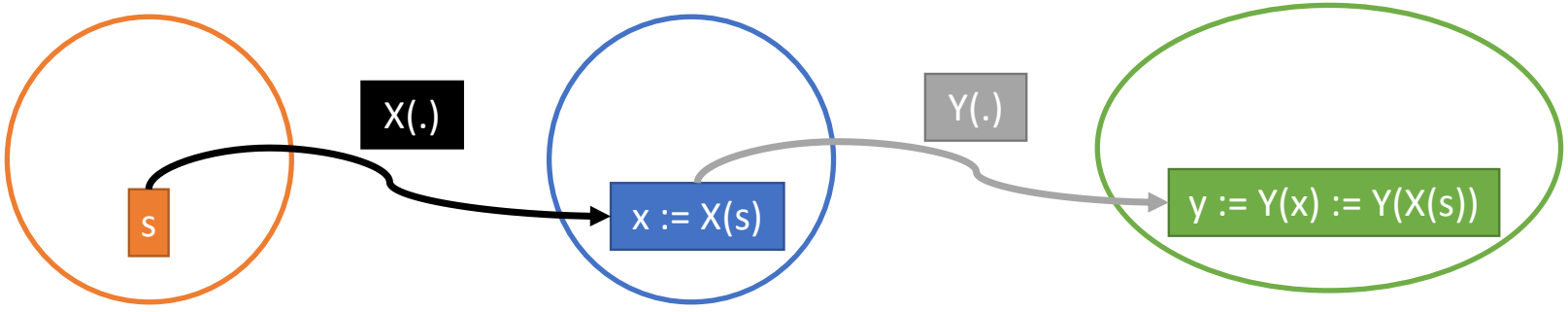
$$E[aX] = \sum_{s\in\Omega}aX(s)P(s) = aE[X]$$

  - This generalizes to:

$$E[\sum_i a_i X_i] = \sum_i a_i E[X_i]$$

# Expectation

- **Expectation of a "function of a random variable"**
  - Let us define values y := Y(x), or "Y(.) is a function of the random variable X"




  - **Discrete** random variable: $E[Y(X)] := E_{P(X)}[Y(X)] := \sum_{x_i} Y(x_i)P(x_i)$
  - **Continuous** random variable: $E[Y(X)] := E_{P(X)}[Y(X)] := \int_x Y(x)P(x)dx$
  - Property:
    - Just as $E_{P(S)}[X(S)] = E_{P(X)}[X]$, …
    - … we get $E_{P(X)}[Y(X)] = E_{P(Y)}[Y]$

$$E[X] := \sum_i x_i p(X = x_i) = \sum_i x_i \left( \sum_{s \in \Omega : X(s) = x_i} P(s) \right) = \sum_{s \in \Omega} X(s) P(s)$$

# Expectation

- Expectation of a function of multiple random variables
  - **Definition:** When we have multiple random variables $X_1,\ldots,X_n$ with a joint PMF/PDF $P(X_1,\ldots,X_n)$ and a function of the multiple random variables $g(X_1,\ldots,X_n)$, then we define the expectation of $g(X_1,\ldots,X_n)$ as:

$$E[g(X_1,\ldots,X_n)] = \sum_{x_1,\ldots,x_n} g(x_1,\ldots,x_n)P(X_1 = x_1,\ldots,X_n = x_n)$$

or

$$E[g(X_1,\ldots,X_n)] = \int_{x_1,\ldots,x_n} g(x_1,\ldots,x_n)P(x_1,\ldots,x_n)\,dx_1\ldots dx_n$$

  - **If X and Y are independent, then E[XY] = E[X] E[Y]**
    - Proof:
    - $\sum_{x,y} xyP(X = x, Y = y) = \sum_{x,y} xyP(X = x)P(Y = y) = \sum_x xP(X = x)\sum_y yP(Y = y)$

# Expectation

- **Tail-sum formula**
  - Let X be a **discrete** random variable taking values in set of **natural** numbers
  - Then, $E(X) = \sum_{k=1}^{\infty} \Pr(X \geq k)$
  - Proof: $E(X) = \sum_{x=1}^{\infty} x \Pr(X = x)$

P(x=1)
P(x=2) P(x=2)
P(x=3) P(x=3) P(x=3)
P(x=4) P(x=4) P(x=4) P(x=4)
…

$$= \sum_{x=1}^{\infty} \sum_{k=1}^{x} \Pr(X = x)$$ Sum over rows (row number = x)

$$= \sum_{k=1}^{\infty} \sum_{x=k}^{\infty} \Pr(X = x)$$ Sum over columns (column number = k)

$$= \sum_{k=1}^{\infty} \Pr(X \geq k)$$

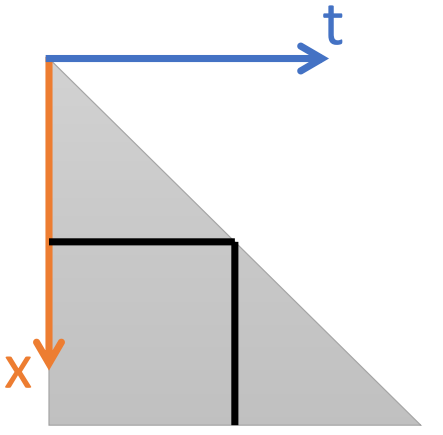
# Expectation

- Tail-sum formula
  - Let X be a **continuous** random variable taking **non-negative** values
    - Notation: For random variable X, PDF is $f_X(.)$ and CDF is $F_X(.)$
  - Then,

$$E(X) = \int_0^\infty (1 - F_X(x))\,\mathrm{d}x$$

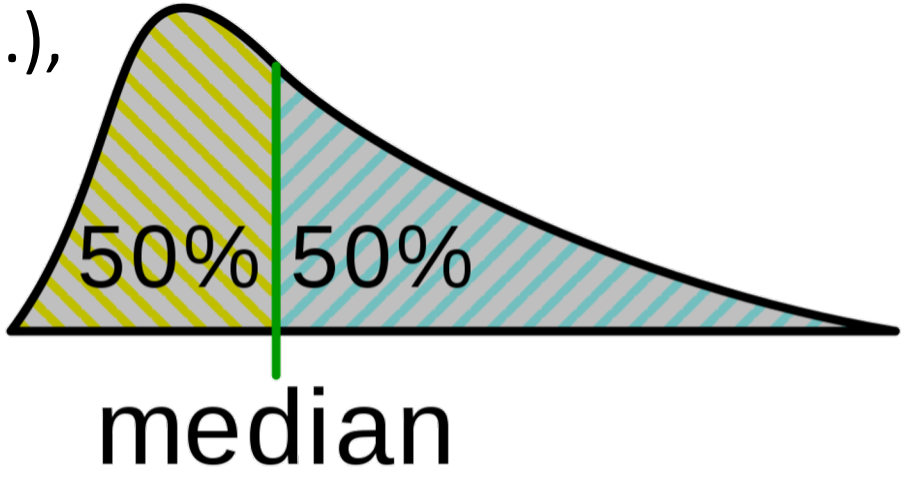
  - Proof:

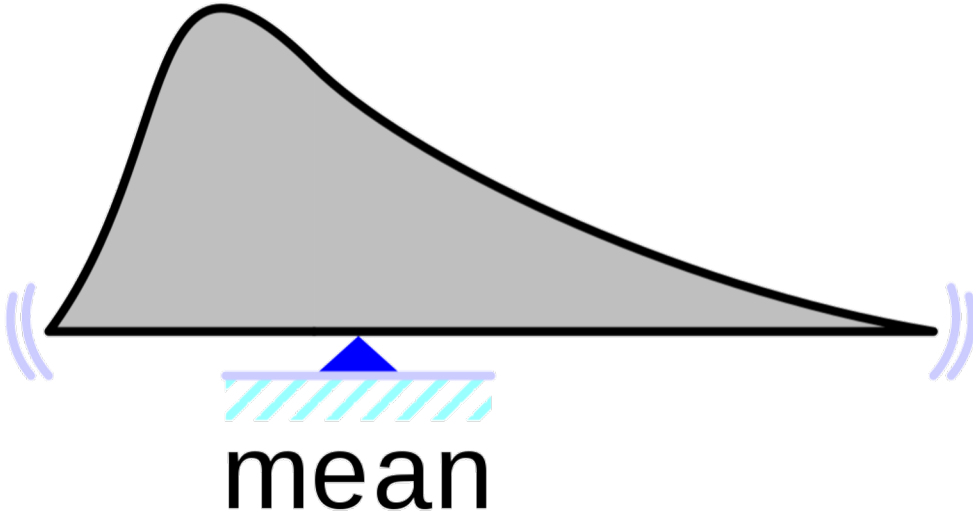
$$\begin{aligned} E(X) &= \int_0^\infty x f_X(x)\,\mathrm{d}x \\ &= \int_0^\infty \int_0^x f_X(x)\,\mathrm{d}t\,\mathrm{d}x \\ &= \int_0^\infty \int_t^\infty f_X(x)\,\mathrm{d}x\,\mathrm{d}t \\ &= \int_0^\infty \Pr(X > t)\,\mathrm{d}t \\ &= \int_0^\infty (1 - F_X(t))\,\mathrm{d}t \end{aligned}$$



# Median

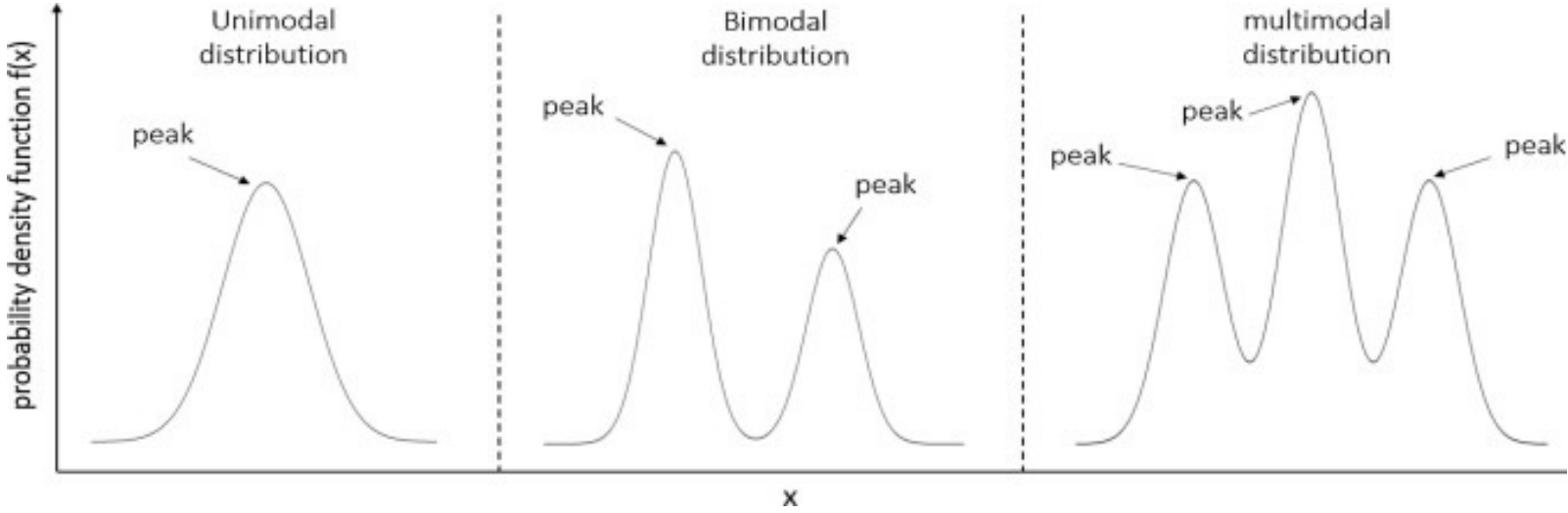
- Definition:
  For a random variable with a PDF/PMF P(.), the median is any real number ‘m’ such that $P(X \leq m) = P(X > m)$
  - Half the probability mass is on the left and half on the right
  - A PDF/PMF can have multiple medians



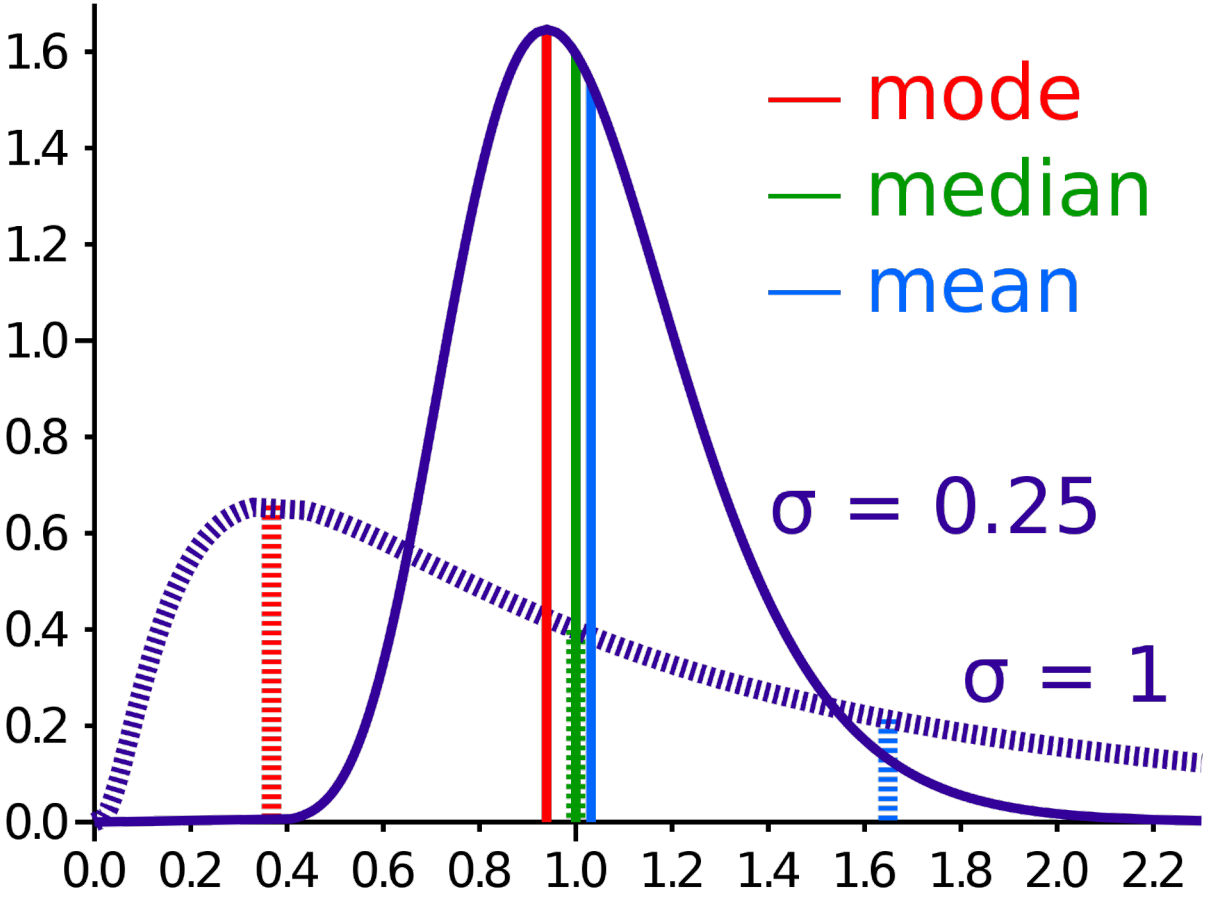
# Mode

- For **discrete** X
  - Mode m is a value for which the PMF value P(X=m) is maximum
  - A PMF can have multiple modes
- For **continuous** X
  - Mode ‘m’ is any **local** maximum of the PDF P(.)
  - A PDF can have multiple modes
  - **Unimodal** PDF = A PDF having only 1 local maximum
  - **Bimodal** PDF: 2 local maxima
  - **Multimodal** PDF: 2 or more local maxima



# Mean, Median, Mode

- For unimodal and symmetric distributions (e.g., Gaussian PDF), mode = mean = median
  - Assuming symmetry around mode, mass on left of mode = mass on right of mode
    - So, mode = median
  - Assuming symmetry around mode, every P(x)dx mass on left of mode is matched by a P(x)dx mass on right of mode
    - So, mode = mean



# Variance

- Measure of the spread of the mass (in PMF or PDF) around the mean
- **Definition**: $\mathrm{Var}(X) := E[(X-E[X])^2]$
- Property: Variance is always non-negative
- Property: $\mathrm{Var}(X) = E[X^2] - (E[X])^2$
  - Proof: LHS =

$E[(X-E[X])^2]$

$= E[\ X^2 + (E[X])^2 - 2.X.E[X]\ ]$

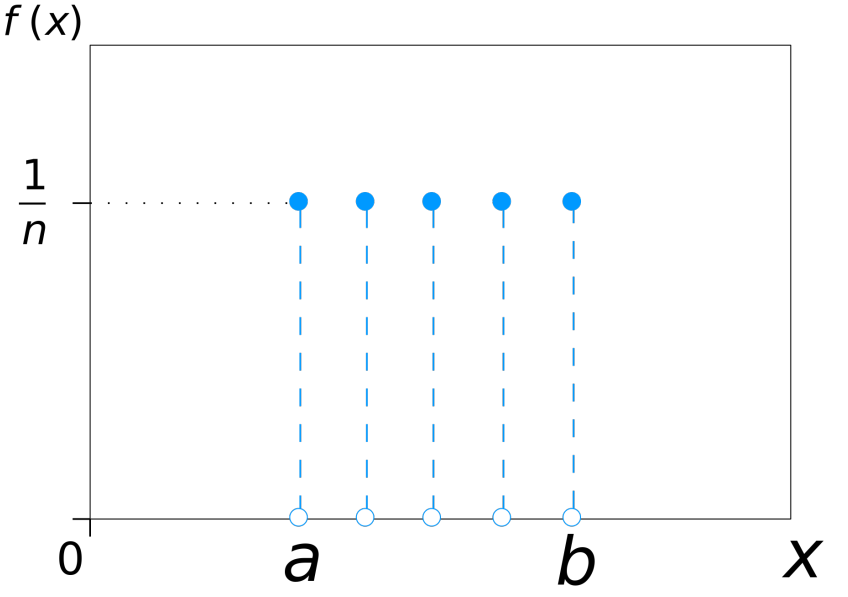
$= E[X^2] + (E[X])^2 - \ 2(E[X])^2$

$= E[X^2] - (E[X])^2 = \mathrm{RHS}$

- **Definition**: **Standard deviation** is the square root of the variance
- Units of variance = square of units of values taken by random variable
- Units of standard deviation = units of values taken by random variable

# Variance

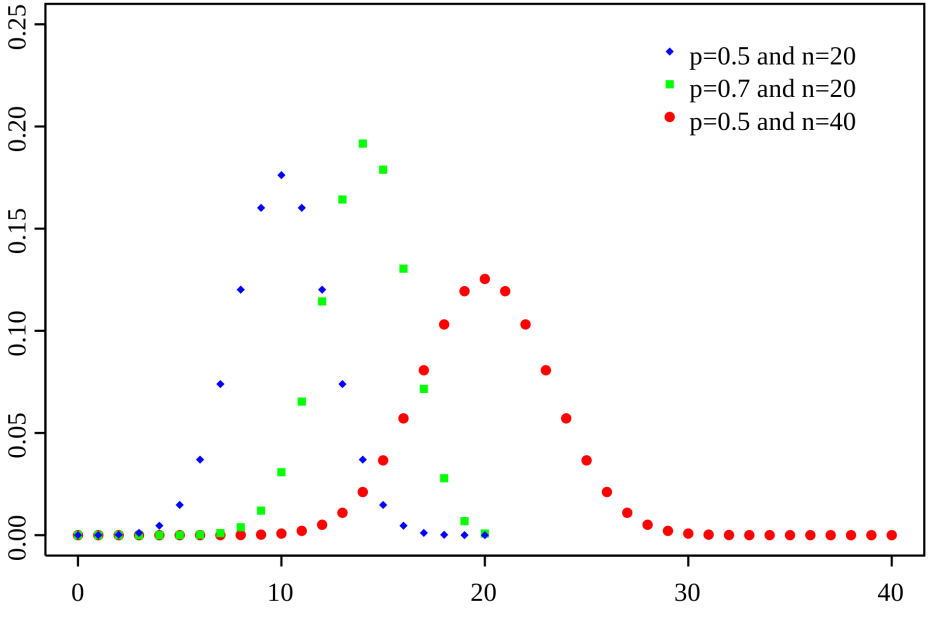
- **Variance of a Uniform Random Variable**
  - Discrete case
    - X has uniform distribution over n integers {a, a+1, …, b}
    - Here, n = b–a+1
    - Variance = $(n^2 - 1) / 12$



# Variance

- **Variance of a Binomial Random Variable**
  - Var(X) = E[X²] – (E[X])² , where E[X] = np

$$
\begin{aligned}
\mathrm{E}\left(X^2\right) &= \sum_{k\geq 0}^{n} k^2 \binom{n}{k} p^k q^{n-k} \\
&= \sum_{k=0}^{n} kn \binom{n-1}{k-1} p^k q^{n-k} \\
&= np \sum_{k=1}^{n} k \binom{n-1}{k-1} p^{k-1} q^{(n-1)-(k-1)} \\
&= np \sum_{j=0}^{m} (j+1) \binom{m}{j} p^j q^{m-j} \\
&= np \left( \sum_{j=0}^{m} j \binom{m}{j} p^j q^{m-j} + \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \right)
\end{aligned}
$$




$$j = k - 1, m = n - 1$$

# Variance

- Variance of a Binomial Random Variable
  - $Var(X) = E[X^2] - (E[X])^2$ , where $E[X] = np$
  - So, $E[X^2]$

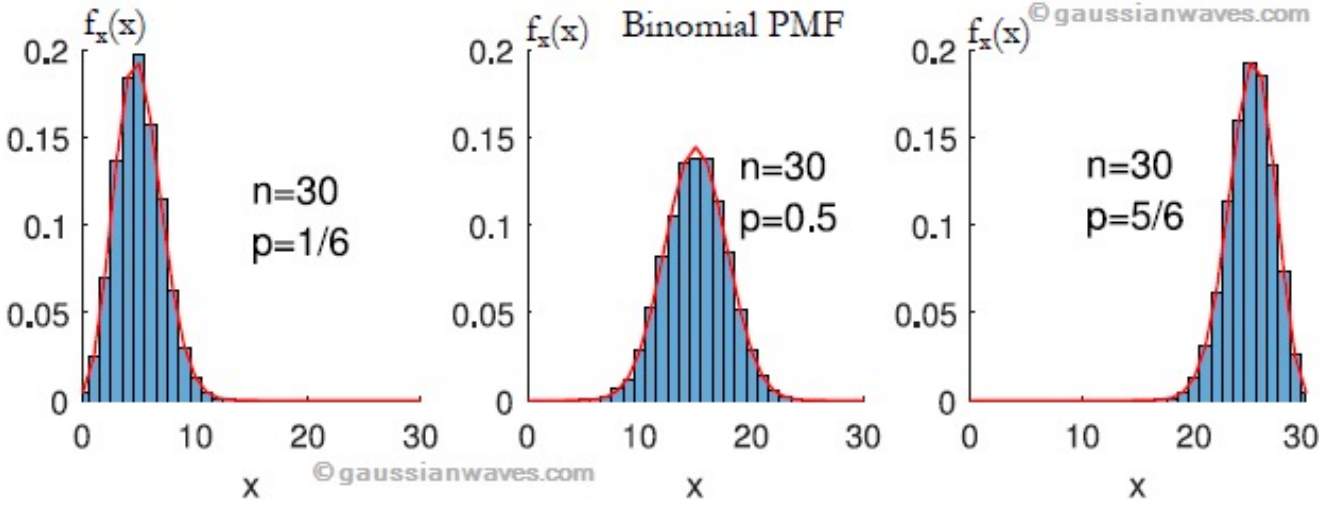
$$= np\left(\sum_{j=0}^{m} j\binom{m}{j}p^j q^{m-j} + \sum_{j=0}^{m}\binom{m}{j}p^j q^{m-j}\right)$$

  = $np\,(mp + 1)$

  = $np\,((n-1)p + 1)$

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

  - Thus, $Var(X) = np(1-p) = npq$
  - Interpretation
    - When p=0 or p=1, then Var(X) = 0, which is the minimum possible
    - When p=q=0.5, then Var(X) is maximized

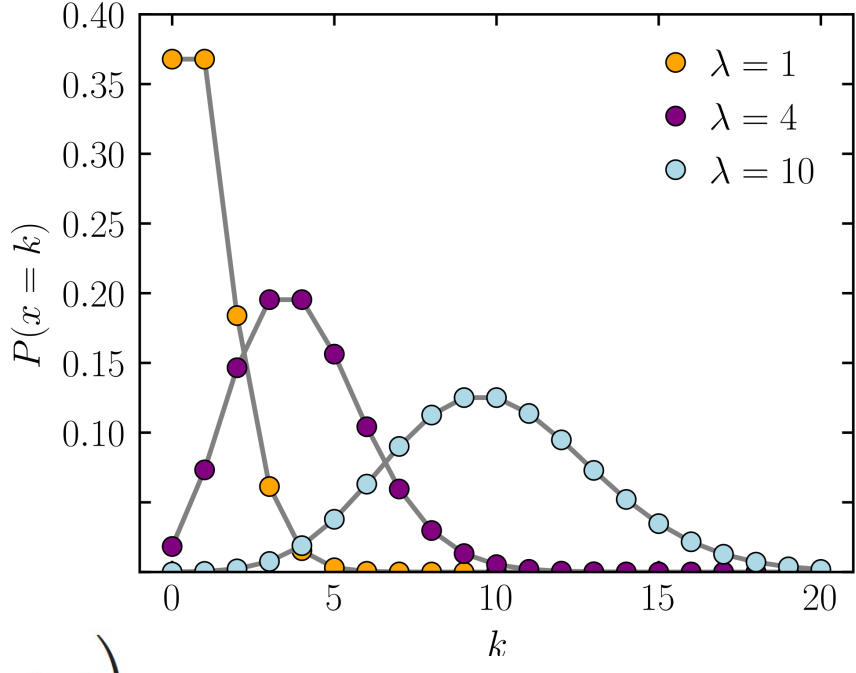


# Variance

- **Variance of a Poisson Random Variable**
  - Var(X) = E[X²] – (E[X])² , where E[X] = λ

$$\begin{aligned}
\mathrm{E}\left(X^2\right) &= \sum_{k\geq 0} k^2 \frac{1}{k!}\lambda^k e^{-\lambda} \\
&= \lambda e^{-\lambda} \sum_{k\geq 1} k \frac{1}{(k-1)!}\lambda^{k-1} \\
&= \lambda e^{-\lambda} \left( \sum_{k\geq 1} (k-1) \frac{1}{(k-1)!}\lambda^{k-1} + \sum_{k\geq 1} \frac{1}{(k-1)!}\lambda^{k-1} \right) \\
&= \lambda e^{-\lambda} \left( \lambda \sum_{k\geq 2} \frac{1}{(k-2)!}\lambda^{k-2} + \sum_{k\geq 1} \frac{1}{(k-1)!}\lambda^{k-1} \right) \\
&= \lambda e^{-\lambda} \left( \lambda \sum_{i\geq 0} \frac{1}{i!}\lambda^{i} + \sum_{j\geq 0} \frac{1}{j!}\lambda^{j} \right)
\end{aligned}$$

$i = k-2, j = k-1$



# Variance

- Variance of a Poisson Random Variable
  - $Var(X) = E[X^2] - (E[X])^2$ , where $E[X] = \lambda$
  - So, $E[X^2]$

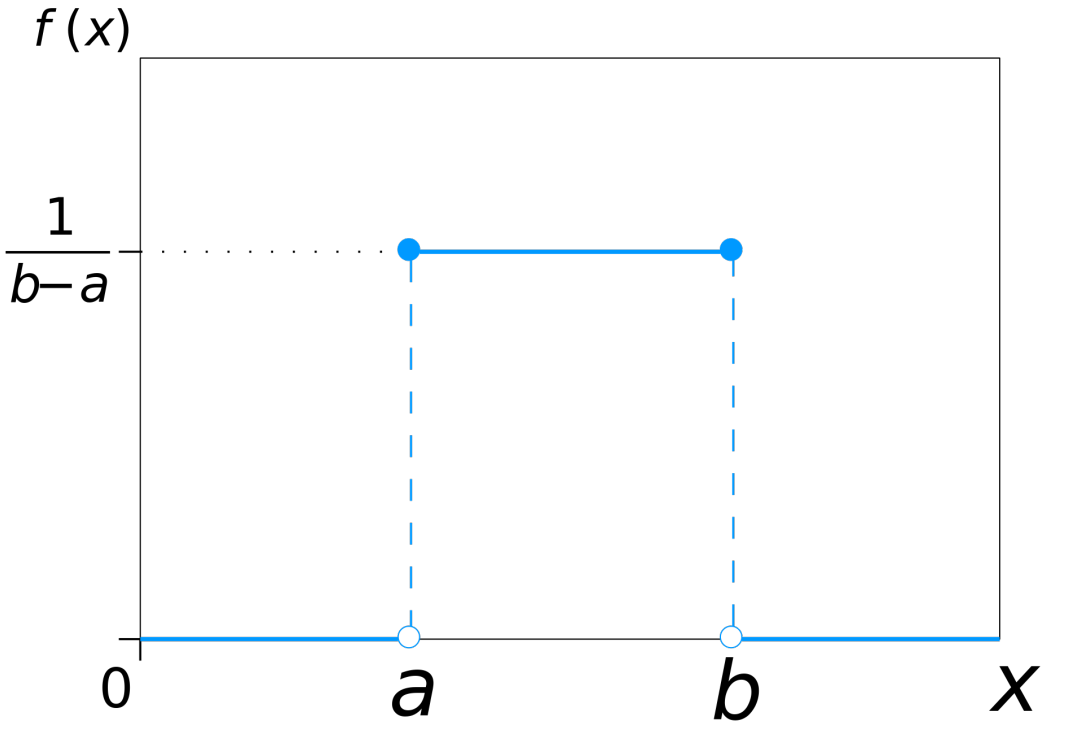
    $= \lambda (\lambda.1 + 1)$

    $= \lambda^2 + \lambda$
  - Thus, $Var(X) = \lambda$
  - Interpretation
    - Mean of Poisson random variable was also $\lambda$
    - Standard deviation of Poisson random variable is $\lambda^{0.5}$
    - As mean increases, so does variance (and standard deviation)
    - When mean increase by factor of N (i.e., N time larger signal = number of arrivals/hits), then the standard deviation (spread) increases only by a factor of $N^{0.5}$
    - As N increases,
      then variability in number of arrivals/hits, relative to average arrival/hit rate, decreases

$$= \lambda e^{-\lambda} \left( \lambda \sum_{i \geq 0} \frac{1}{i!} \lambda^i + \sum_{j \geq 0} \frac{1}{j!} \lambda^j \right)$$

# Variance

- **Variance of a Uniform Random Variable**
  - Continuous case
    - X has uniform distribution over [a,b]
    - Variance = $(b - a)^2 / 12$



# Variance

**PDF**

P(x) = 0, for all x < 0

P(x) = $\lambda \exp(-\lambda x), \forall x \geq 0$

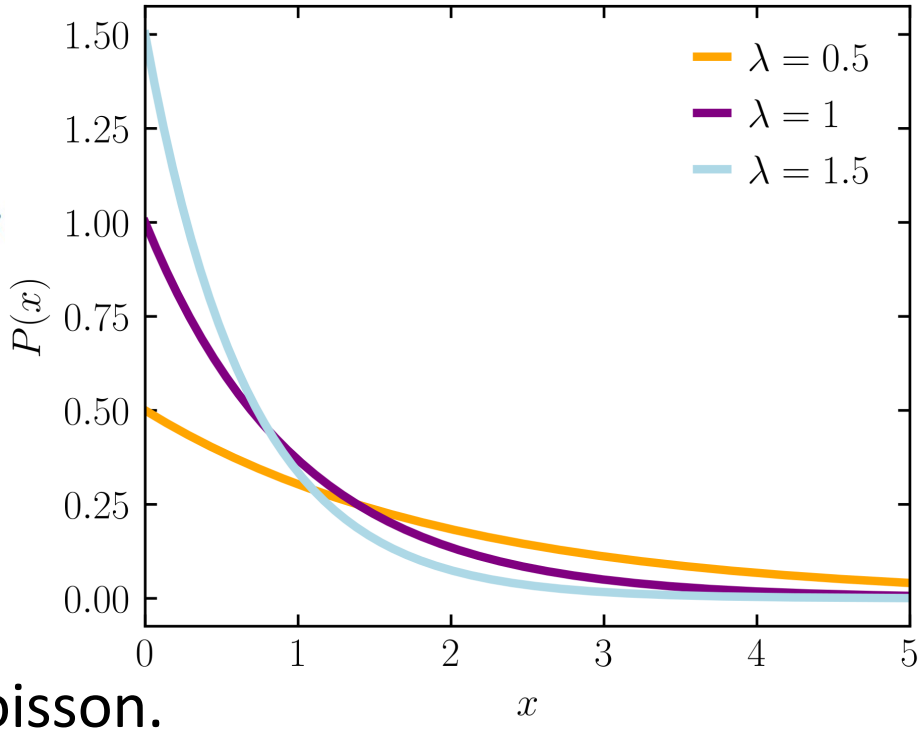
**CDF**

f(x) = 0, for all x < 0

f(x) = $1 - \exp(-\lambda x), \forall x \geq 0$

- **Variance of a Exponential Random Variable**
  - Var(X) = $E[X^2] - (E[X])^2$ , where $E[X] = \beta := 1/\lambda$

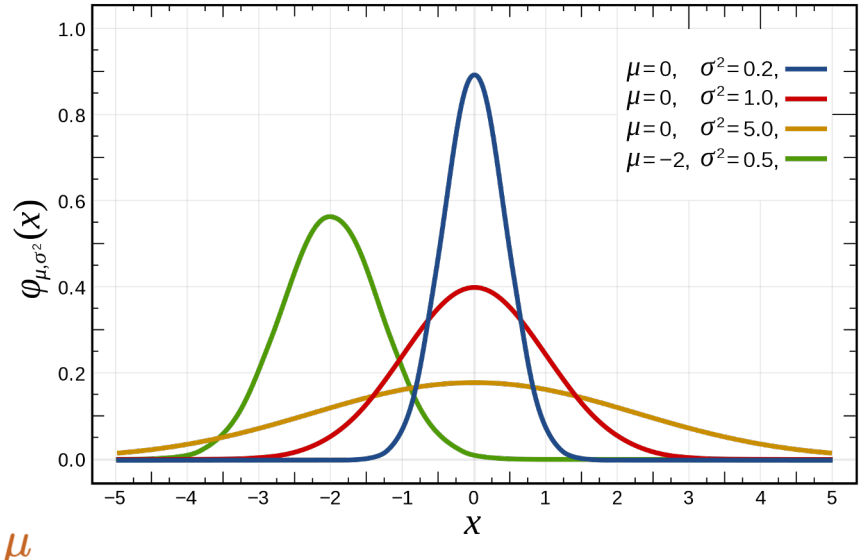
$$
\begin{aligned}
\mathrm{E}\left(X^2\right) &= \int_{x \in \Omega X} x^2 f_X(x)\, \mathrm{d}x \\
&= \int_0^\infty x^2 \frac{1}{\beta} \exp\left(-\frac{x}{\beta}\right) \mathrm{d}x \\
&= \left[-x^2 \exp\left(-\frac{x}{\beta}\right)\right]_0^\infty + \int_0^\infty 2x \exp\left(-\frac{x}{\beta}\right) \mathrm{d}x \\
&= 0 + 2\beta \int_0^\infty x \frac{1}{\beta} \exp\left(-\frac{x}{\beta}\right) \mathrm{d}x \\
&= 2\beta\, \mathrm{E}\left(X\right) \\
&= 2\beta^2
\end{aligned}
$$




- So, Var(X) = $\beta^2$ . So, $\beta = E[X] = SD(X)$; unlike Poisson.

# Variance

- **Variance of a Gaussian Random Variable**
  - Var(X) = E[X²] – (E[X])², where E[X] = μ




$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x - \mu^2$$

$$= \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sqrt{2}\sigma t + \mu)^2 \exp\left(-t^2\right) \mathrm{d}t - \mu^2 \qquad t = \frac{x-\mu}{\sqrt{2}\sigma}$$

$$= \frac{1}{\sqrt{\pi}} \left( 2\sigma^2 \int_{-\infty}^{\infty} t^2 \exp\left(-t^2\right) \mathrm{d}t + 2\sqrt{2}\sigma\mu \int_{-\infty}^{\infty} t \exp\left(-t^2\right) \mathrm{d}t + \mu^2 \int_{-\infty}^{\infty} \exp\left(-t^2\right) \mathrm{d}t \right) - \mu^2$$

$$= \frac{1}{\sqrt{\pi}} \left( 2\sigma^2 \int_{-\infty}^{\infty} t^2 \exp\left(-t^2\right) \mathrm{d}t + 2\sqrt{2}\sigma\mu \cdot 0 \right) + \mu^2 - \mu^2$$

# Variance

- Variance of a Gaussian Random Variable
  - Var(X) = $E[X^2] - (E[X])^2$ , where $E[X] = \mu$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^2 \exp\left(-t^2\right) \mathrm{d}t$$

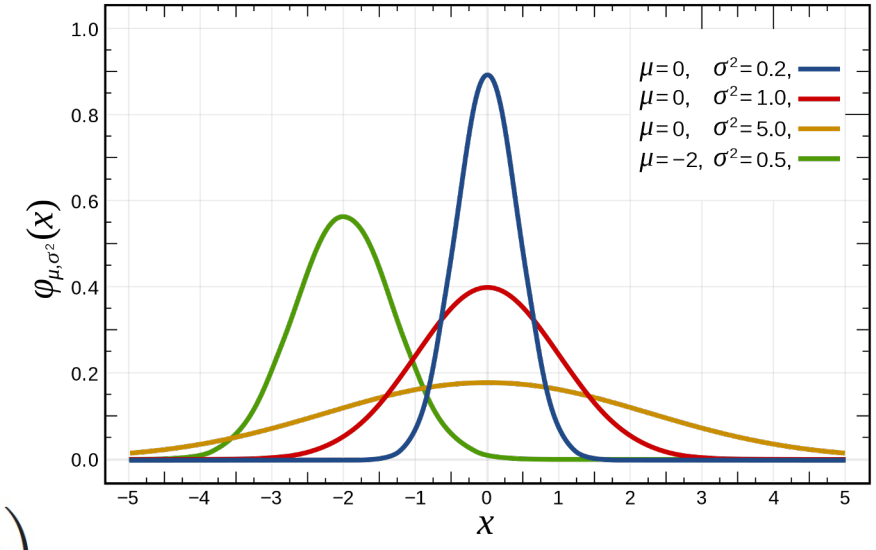
t . (t.exp(-$t^2$))

$$= \frac{2\sigma^2}{\sqrt{\pi}} \left( \left[ -\frac{t}{2} \exp\left(-t^2\right) \right]_{-\infty}^{\infty} + \frac{1}{2} \int_{-\infty}^{\infty} \exp\left(-t^2\right) \mathrm{d}t \right)$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \cdot \frac{1}{2} \int_{-\infty}^{\infty} \exp\left(-t^2\right) \mathrm{d}t$$

$$= \frac{2\sigma^2 \sqrt{\pi}}{2\sqrt{\pi}}$$

$$= \sigma^2$$



# Variance

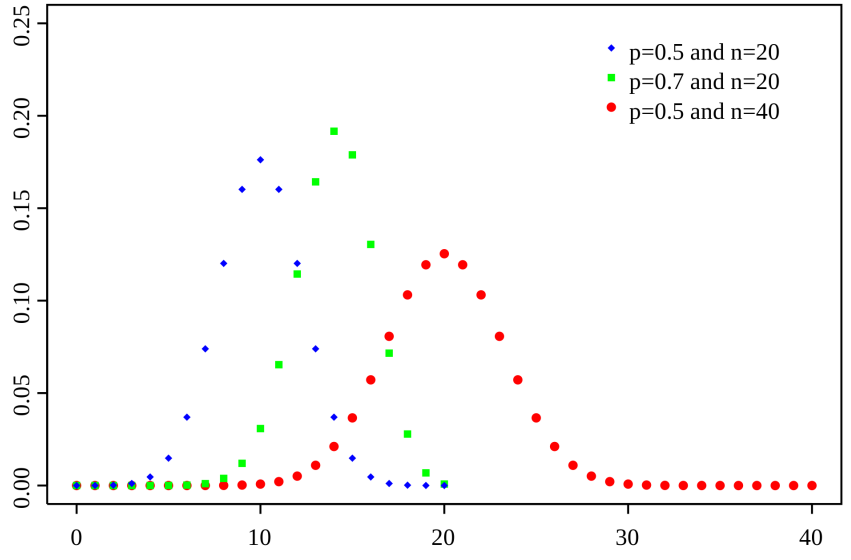
- Example
  - **Variance of a limiting case of binomial**
  - As n tends to infinity,

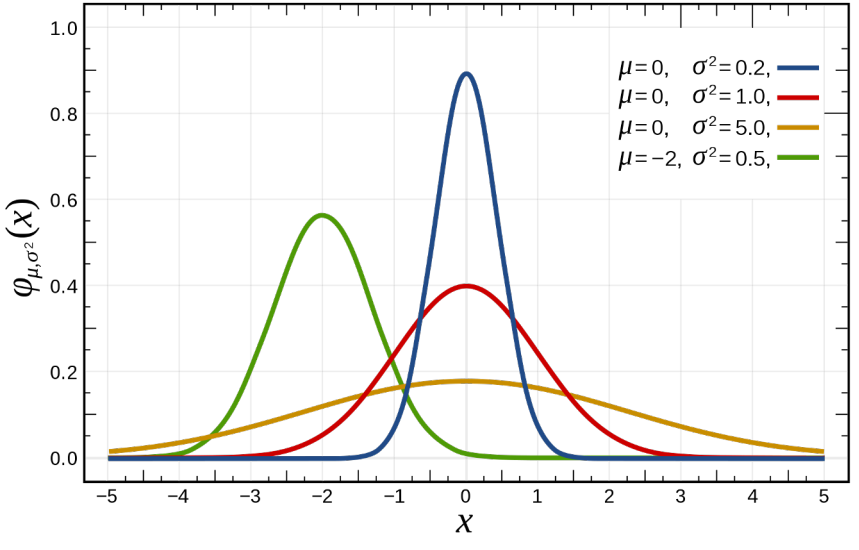
binomial $$f(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

tends to

Gaussian $$\frac{1}{\sqrt{2\pi npq}} e^{-(x-np)^2/2npq}$$

  - Gaussian variance $\sigma^2$ (= npq in this case) is consistent with binomial variance npq



# Variance

- **Property: Var(aX+c) = $a^2$Var(X)**
  - Adding a constant to a random variable doesn't change the variance (spread)
    - This only shifts the PDF/PMF
    - If Y := X + c, then Var(Y) = Var(X)
  - If we scaling a random variable by 'a', then the variance gets scaled by $a^2$
    - If Y := aX, then Var(Y) = $a^2$Var(X)
  - Proof:

$$\begin{aligned}
\operatorname{Var}(aX+c) &= E((aX+c)^2) - E(aX+c)^2 \\
&= E(a^2X^2 + 2acX + c^2) - (aE(X)+c)^2 \\
&= a^2E(X^2) + 2acE(X) + c^2 - a^2E(X)^2 - 2acE(X) - c^2 \\
&= a^2(E(X^2) - E(X)^2) \\
&= a^2 \operatorname{Var}(X)
\end{aligned}$$

# Variance

- **Property: Var(X+Y) = Var(X) + Var(Y) + 2(E[XY] – E[X]E[Y])**
- Proof:

$$\begin{aligned}
\mathrm{Var}(X+Y) &= E((X+Y)^2) - E(X+Y)^2 \\
&= E(X^2 + 2XY + Y^2) - (E(X) + E(Y))^2 \\
&= E(X^2) + 2E(XY) + E[Y^2] - E(X)^2 - 2E(X)E(Y) - E(Y)^2 \\
&= (E(X^2) - E(X)^2) + (E[Y^2] - E(Y)^2) + 2(E(XY) - 2E(X)E(Y)) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2(E(XY) - E(X)E(Y))
\end{aligned}$$

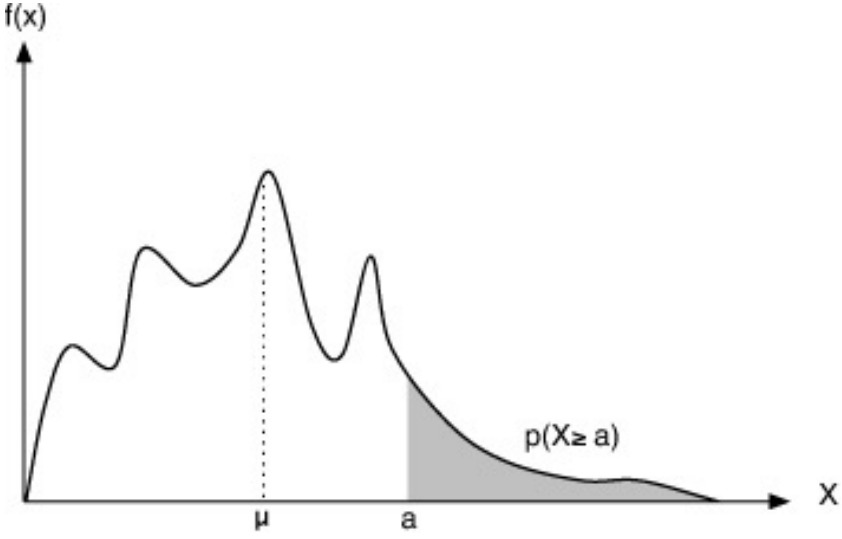
- If X and Y are independent,
  then E[XY] = E[X] E[Y], and so Var(X+Y) = Var(X) + Var(Y)
- If X,Y,Z are independent, then
  Var(X+Y+Z) = Var(X+Y) + Var(Z) = Var(X) + Var(Y) + Var(Z)
- For independent random variables $X_1$, …, $X_n$;
  $\mathrm{Var}(X_1 + \ldots + X_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)$

# Markov's Inequality

- **Theorem**: Let X be a random variable with PDF P(.).
  Let u(.) be an non-negative-valued function.
  Let 'c' be a positive constant.
  Then, **$P(u(X) \geq c) \leq E[u(X)] / c$**
- Proof:
  - $E[u(X)] = \int_{x:u(x)\geq c} u(x)\, P(x)\, dx + \int_{x:u(x)<c} u(x)\, P(x)\, dx$
  - Because u(.) takes non-negative values, each integral above is non-negative
  - So, $E[u(X)] \geq \int_{x:u(x)\geq c} u(x)\, P(x)\, dx$

  $\geq c \int_{x:u(x)\geq c} P(x)\, dx$

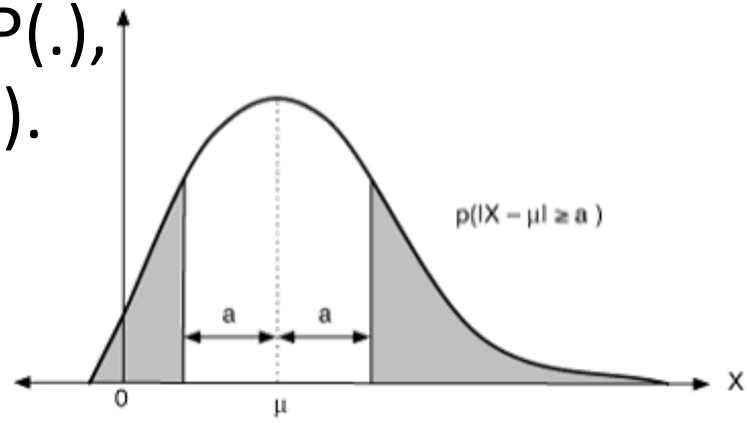
  $= c\, P(u(X) \geq c)$
  - Because c>0, we get $E[u(X)]/c \geq P(u(X) \geq c)$
- Special case →
  - X is non-negative & u(x) := x



# Chebyshev's Inequality

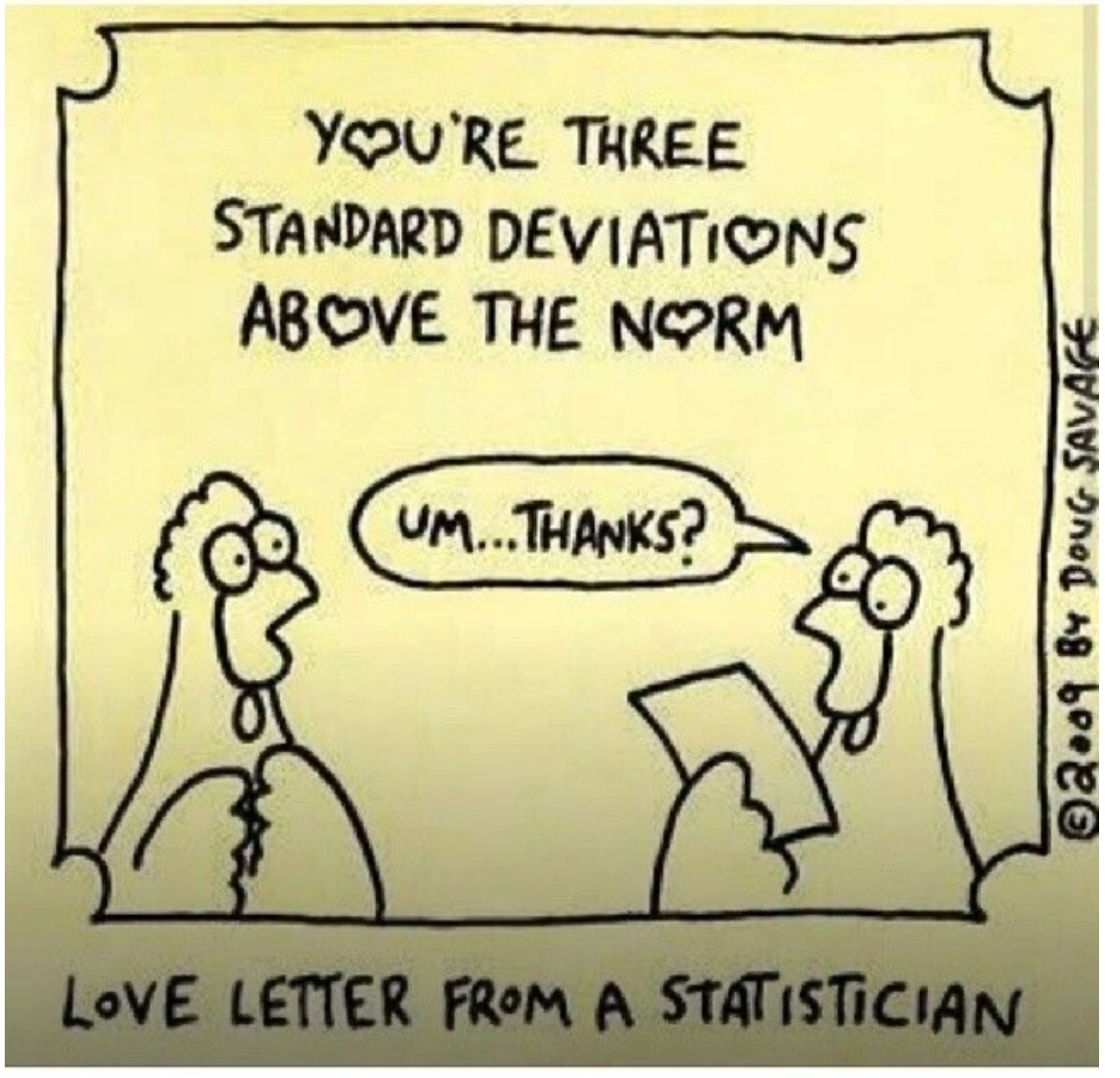
Markov's Inequality:
$P(u(X) \geq c) \leq E[u(X)] / c$

- **Theorem**: Let X be a random variable with PDF P(.), finite expectation E[X], and finite variance Var(X). Then, **$P(|X-E[X]| \geq a) \leq Var(X) / a^2$**




- Proof:
  - Define random variable $u(X) := (X-E[X])^2$
  - Then, by Markov's inequality, $P((X-E[X])^2 \geq a^2) \leq E[(X-E[X])^2] / a^2$
  - LHS = $P(|X-E[X]| \geq a)$
  - RHS = $Var(X) / a^2$
  - Q.E.D.
- **Corollary**: If random variable X has standard deviation σ, then $P(|X-E[X]| \geq k\sigma) \leq 1/k^2$
  - This is consistent with the notion of standard deviation (σ) or variance ($\sigma^2$) measuring the spread of the PDF around the mean (center of mass)

# Chebyshev's Inequality



# Chebyshev

- Pafnuty Chebyshev
  - Founding father of Russian mathematics
  - Students: Lyapunov, Markov
  - First person to think systematically in terms of random variables and their moments and expectations

# Markov

- Andrey Markov
  - Russian mathematician best known for his work on stochastic processes
  - Advisor: Chebyshev
  - Students: Voronoy
  - One year after doctoral defense, appointed extraordinary professor
  - He figured out that he could use chains to model the alliteration of vowels and consonants in Russian literature

А. А. Марков (1886).

# Jensen's Inequality

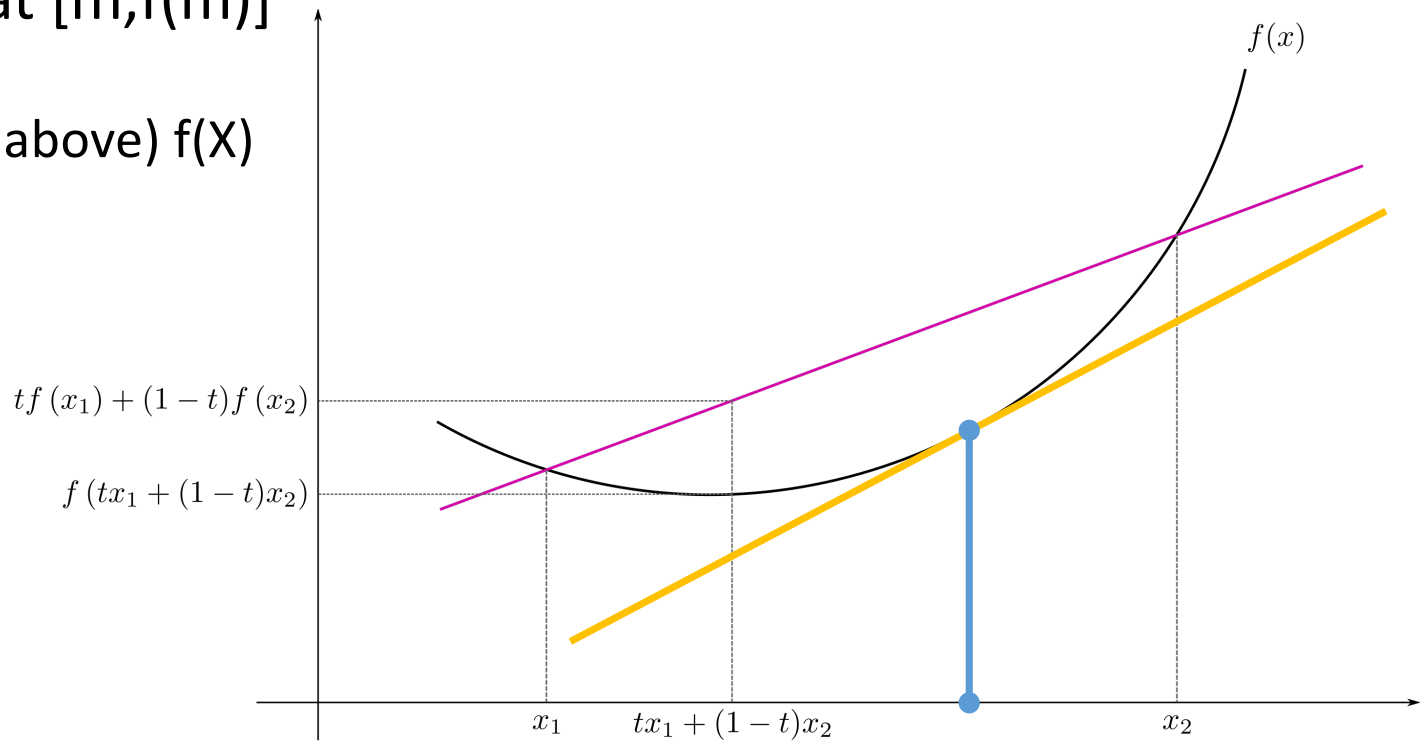
- **Theorem**: Let X be a random variable. Let f(.) be a **convex** function. Then, **E[f(X)] ≥ f(E[X])**

A real-valued function is called convex if the line segment between any two points on the graph of the function lies above the graph between the two points

- Proof:
  - Let m := E[X]
  - Consider a tangent to f(.) at [m,f(m)]
    - This line is, say, Y = aX+b, which lies at/below (never above) f(X)
    - Then, f(m) = am+b
  - Then, E[f(X)] ≥ E[aX+b]

  = aE[X] + b

  = f(E[X])

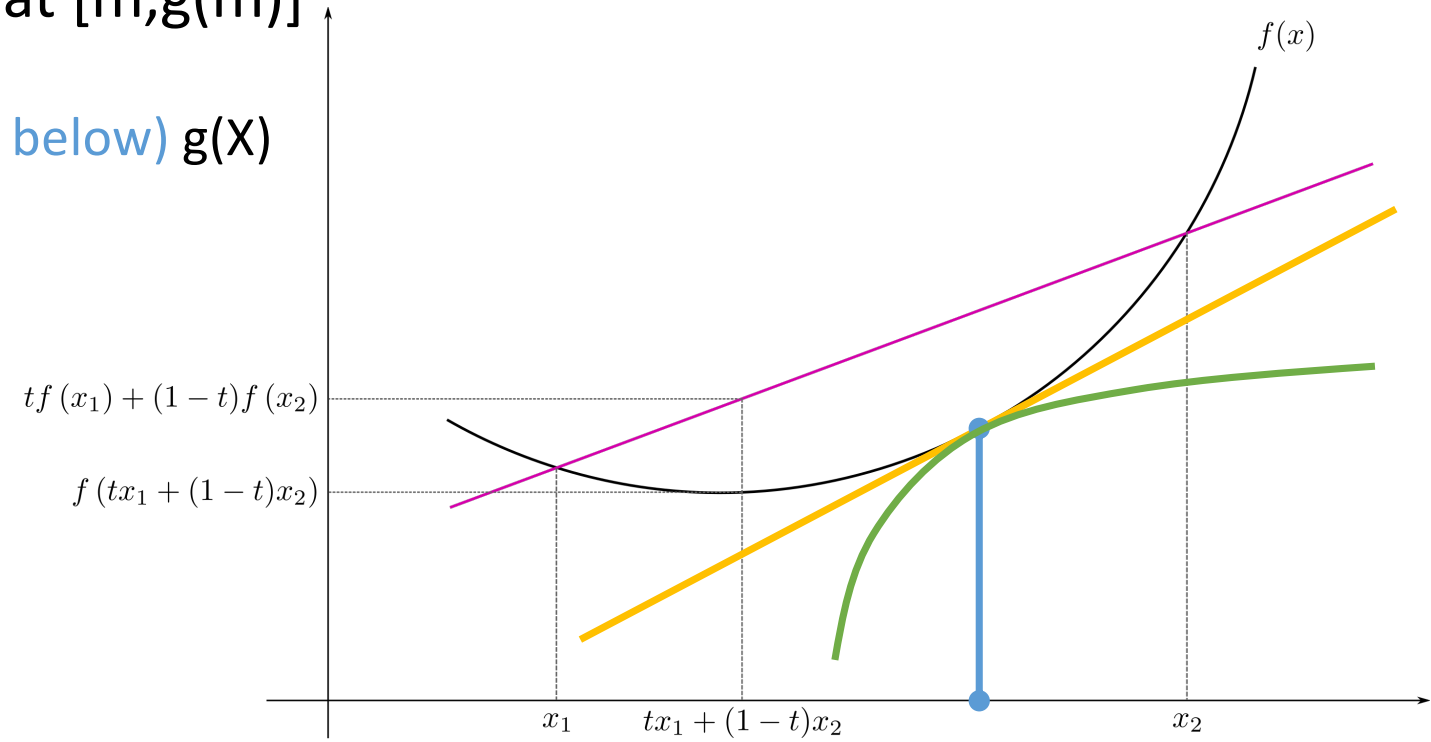


# Jensen's Inequality

- **Corollary**: Let X be a random variable. Let g(.) be a concave function. Then, **E[g(X)] ≤ g(E[X])**
- Proof:
  - Let m := E[X]
  - Consider a tangent to g(.) at [m,g(m)]
    - This line is, say, Y = cX+d, which lies at/above (never below) g(X)
    - Then, g(m) = am+b
  - Then, E[g(X)] ≤ E[aX+b]

  = aE[X] + b

  = g(E[X])



# Jensen

- Johan Jensen
  - Danish mathematician and engineer
  - President of the Danish Mathematical Society from 1892 to 1903
  - Never held any academic position
  - Engineer for Copenhagen Telephone Company
    - Became head of its technical department
  - Learned advanced math topics by himself
  - All his mathematics research was carried out in his spare time

# Minimizer of Expected Absolute Deviation

- **Theorem**: E[|X – c|] is minimum when c = Median(X)
- Case 1: c ≤ Median(X) = m
  - **E[|X – c|]** = $\int_{-\infty}^{c}(c-x)P(x)dx + \int_{c}^{\infty}(x-c)P(x)dx$ (say, **A + B**)
  - A = $\int_{-\infty}^{m}(c-x)P(x)dx - \int_{c}^{m}(c-x)P(x)dx$ (say, A1 – A2)
  - B = $\int_{c}^{m}(x-c)P(x)dx + \int_{m}^{\infty}(x-c)P(x)dx$ (say, B1 + B2)
  - Now, B1 – A2 = $2\int_{c}^{m}(x-c)P(x)dx \geq 0$
  - A1 = $\int_{-\infty}^{m}(c-m)P(x)dx + \int_{-\infty}^{m}(m-x)P(x)dx$ (say, A11 + A12)
    - A11 = (c – m) 0.5 (by definition of median)
  - B2 = $\int_{m}^{\infty}(x-m)P(x)dx + \int_{m}^{\infty}(m-c)P(x)dx$ (say, B21 + B22)
    - B22 = (m – c) 0.5 (by definition of median)
  - So, A1 + B2 = A12 + B21 = E[|X – m|]
  - Value of c minimizing A+B is c = m

# Minimizer of Expected Absolute Deviation

- **Theorem**: E[|X – c|] is minimum when c = Median(X)
- Case 2: Median(X) = m ≤ c
  - **E[|X – c|]** = $\int_{-\infty}^{c}(c-x)P(x)dx + \int_{c}^{\infty}(x-c)P(x)dx$ (say, **A + B**)
  - A = $\int_{-\infty}^{m}(c-x)P(x)dx + \int_{m}^{c}(c-x)P(x)dx$ (say, A1 + A2)
  - B = $-\int_{m}^{c}(x-c)P(x)dx + \int_{m}^{\infty}(x-c)P(x)dx$ (say, – B1 + B2)
  - Now, A2 – B1 = $2\int_{m}^{c}(c-x)P(x)dx \geq 0$
  - As before, A1 + B2 = A12 + B21 = E[|X – m|]
  - Value of c minimizing A+B is c = m

# Mean, Median, Standard Deviation

- **Theorem**:
  Mean(X) and Median(X) are within a distance of SD(X) of each other
- Proof:
  - Distance between mean and median

  $= |E[X] - \text{Median}(X)|$

  $= |E[X - \text{Median}(X)]|$

  This is $|E[.]|$, where $|.|$ is a convex function. Apply Jensen's inequality.

  $\leq E[|X - \text{Median}(X)|]$

  $\leq E[|X - E[X]|]$ (because Median(X) minimizes expected absolute deviation)

  $= E[\text{SQRT}\{ (X - E[X])^2 \}]$

  This is E[SQRT(.)], where SQRT(.) is a concave function. Jensen's inequality →

  $\leq \text{SQRT}\{ E[ (X - E[X])^2 ] \}$

  $= \text{SQRT}\{ \text{Var}(X) \} = \text{SD}(X)$

# Law of Large Numbers

- This justifies why the expectation is motivated as an average over a large number of random experiments ("long-term average")
- Let random variables $X_1$, …, $X_i$, …, $X_n$ be 'n' **independent and identically distributed (i.i.d.)**, each with mean $\mu=E[X_i]$ and finite variance $v=Var(X_i)$
- Let the **average**, over 'n' experiments, be modeled by a random variable $\overline{X} := (X_1 + \ldots + X_n) / n$
- Then, the **expected average** $E[\overline{X}] = \mu$, by the linearity of expectation
- But, in specific runs, how close is $E[\overline{X}]$ to the expectation $\mu$ ?
- So, we analyze the spread of the average $E[\overline{X}]$ around $\mu$
- $Var(\overline{X}) := Var(X_1/n) + \ldots + Var(X_n/n) = n(v/n^2) = v/n$

# Law of Large Numbers

- This justifies why the expectation is motivated as an average over a large number of random experiments
- **Law of large numbers**: For all $\varepsilon > 0$, as $n \to \infty$, $P(|\overline{X} - \mu| \geq \varepsilon) \to 0$
- Proof:
  - Using Chebyshev's inequality,

$P(|\overline{X} - \mu| \geq \varepsilon)$

$\leq Var(\overline{X}) / \varepsilon^2$

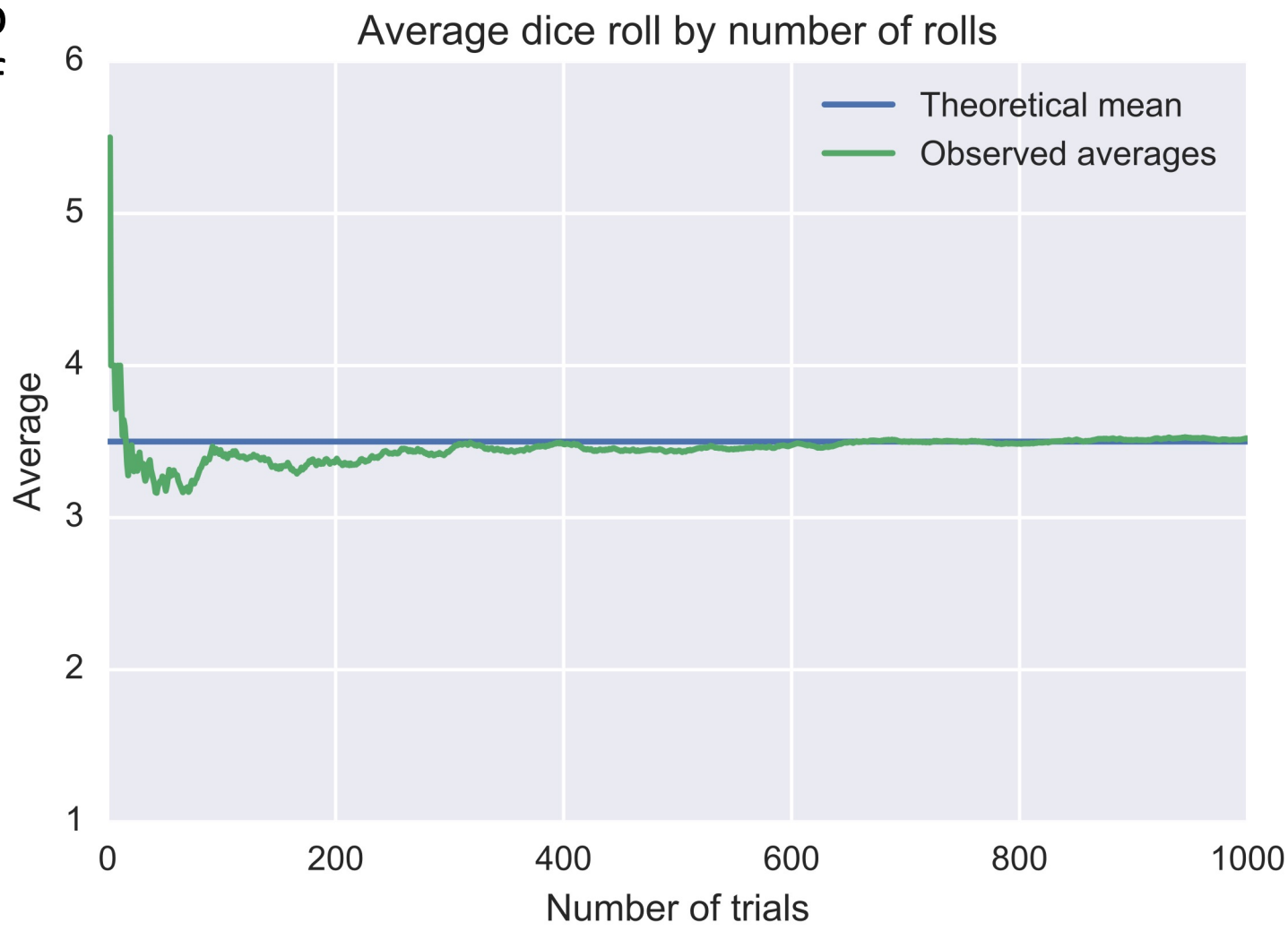
$= v / (n\varepsilon^2)$

$\to 0$, as $n \to \infty$

  - Thus, as the average $\overline{X}$ uses data from more number of experiments 'n', the event of $\overline{X}$ being far from $\mu$ (farther than $\varepsilon$) has a probability that tends to 0

# Law of Large Numbers

- Example
  - This also gives us a way to compute an “estimate” of the **expectation μ of a random variable X** from “observations”/data
    - What is the estimate ?
    - $\overline{X}$

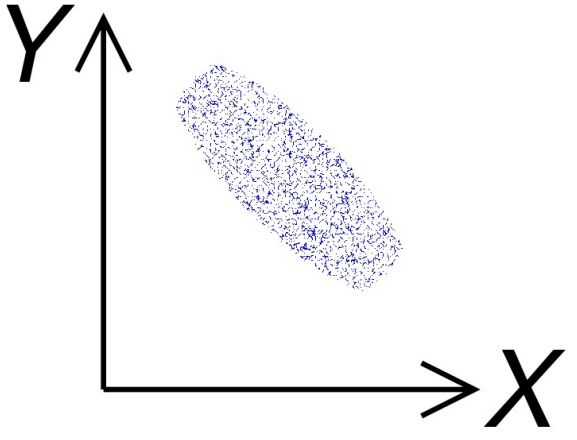


# Covariance

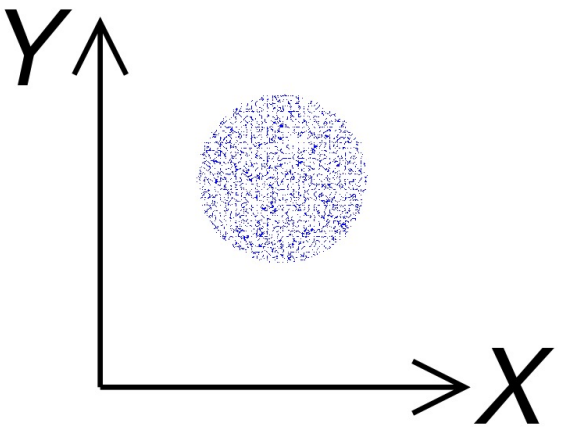
- For random variables X and Y, consider the joint PMF/PDF P(X,Y)
- Covariance: A measure of how the values taken by X and Y vary together ("co"-"vary")
- **Definition**: $\mathrm{Cov}(X,Y) := E[(X - E[X])(Y - E[Y])]$
  - Interpretation:
    - Define $U(X) := X - E[X]$ and $V(Y) := Y - E[Y]$ (Note: U and V have expectation 0)
    - In the joint PDF P(U,V),
      if **larger** (more +ve) values of U typically correspond to **larger** values of V, and
      if smaller (more –ve) values of U typically correspond to smaller values of V,
      then U and V co-vary/correlate **positively**
    - In the joint PDF P(U,V),
      if **larger** values of U typically correspond to **smaller** values of V, and …
      then U and V co-vary/correlate **negatively**
- **Property**: **Symmetry**: $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$

# Covariance

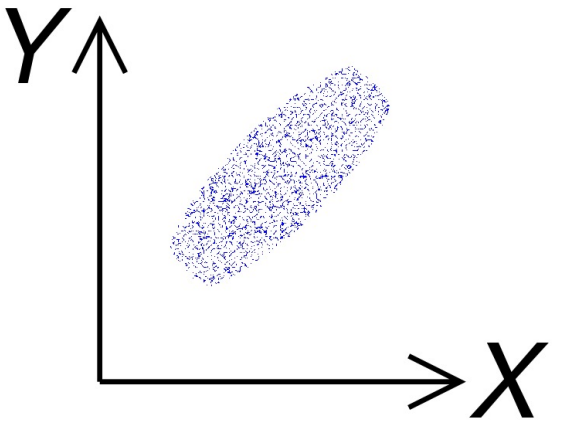
- Examples




$\mathrm{cov}(X,Y) < 0$




$\mathrm{cov}(X,Y) \approx 0$




$\mathrm{cov}(X,Y) > 0$

# Covariance

- **Property**: $Cov(X,Y) = E[XY] - E[X]E[Y]$
  - Proof:
    - $Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$

    $= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y]$

    $= E[XY] - E[X]E[Y]$
  - So, $Var(X+Y) = Var(X) + Var(Y) + 2(E[XY] - E[X]E[Y]) = Var(X) + Var(Y) + 2Cov(X,Y)$
  - Thus, when X and Y are independent, then $Cov(X,Y) = 0$
- **Property**: When $Var(X)=0$ or $Var(Y)=0$, then $Cov(X,Y)=0$
- **Property**: When $Y := mX + c$ (with finite m), what is $Cov(X,Y)$ ?
  - $Cov(X,Y) = E[XY] - E[X]E[Y]$

  $= E[mX^2 + cX] - E[X](m.E[X] + c)$

  $= m.E[X^2] - m(E[X])^2 = m.Var(X)$
  - When $Var(X)>0$, covariance is $\propto$ line-slope ‘m’, and has same sign as that of m

# Covariance

- **Bilinearity of Covariance**
  - Let X, $X_1$, $X_2$, Y, $Y_1$, $Y_2$ be random variables. Let c be a scalar constant.
  - **Property:** $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y) = \text{Cov}(Y, X_1 + X_2)$
    - Proof (first part; second part follows from symmetry):

$$\begin{aligned}\text{Cov}(X_1 + X_2, Y) &= E((X_1 + X_2)Y) - E(X_1 + X_2)E(Y) \\ &= E(X_1Y) - E(X_1)E(Y) + E(X_2Y) - E(X_2)E(Y) \\ &= \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)\end{aligned}$$

  - **Property:** Cov(aX, Y) = a.Cov(X, Y) = Cov(X, aY)
    - Proof (first part):
    - Cov(aX, Y)

    = E[aXY] − E[aX]E[Y]

    = a(E[XY]−E[X]E[Y])

    = a.Cov(X,Y)

# Standardized Random Variable

- **Definition**:
  If X is a random variable, then its **standardized** form is given by:
  X* := (X – E[X]) / SD(X), where SD(.) gives the standard deviation
- **Property**: E[X*] = 0, Var(X*) = 1
  - Proof:

$$E(X^*) = \frac{E(X) - E(X)}{\sigma_X} = 0$$

$$\mathrm{Var}(X^*) = \mathrm{Var}\left(\frac{X - E(X)}{\sigma_X}\right) = \frac{\mathrm{Var}(X)}{\sigma_X^2} = \frac{\mathrm{Var}(X)}{\mathrm{Var}(X)} = 1$$

- X* is unit-less
- X* is obtained by:
  - First shifting/translating X to make mean 0, and
  - Then scaling the shifted variable to make variance 1

# Correlation

- For covariance, the magnitude isn't easy to interpret (unlike its sign)
- **Correlation**: A measure of how the values taken by X and Y vary together ("co"-"relate") obtained by rescaling covariance
  - Pearson's correlation coefficient
  - Assuming X and Y are linearly related, correlation magnitude shows the strength of the (functional/deterministic) relationship between X and Y
- Let 'SD' = standard deviation
- **Definition**: Cor(X,Y) := $E\left[\left(\frac{X-E[X]}{\mathrm{SD}(X)}\right)\left(\frac{Y-E[Y]}{\mathrm{SD}(Y)}\right)\right] = \mathrm{Cov}(X,Y)/(\mathrm{SD}(X)\mathrm{SD}(Y))$

  - Thus, Cor(X,Y) = E[X*Y*], where X* and Y* are the standardized variables

  = E[X*Y*] – E[X*]E[Y*]

  = Cov(X*,Y*)

# Correlation

- **Property**: $-1 \le Cor(X,Y) \le 1$
  - Proof:
    - First inequality
    - $0 \le E[(X^*+Y^*)^2]$

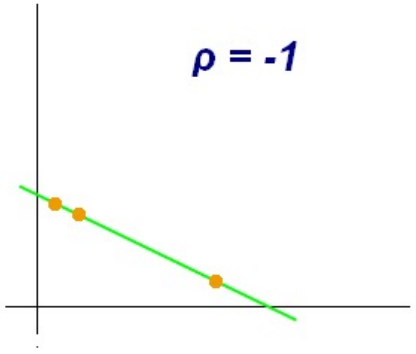
      $= E[(X^*)^2] + E[(Y^*)^2] + 2E[X^*Y^*]$

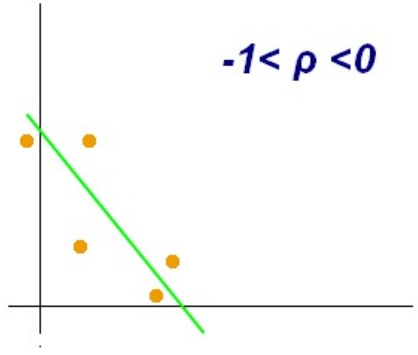
      $= 2(1 + Cor(X,Y))$
    - So, $-1 \le Cor(X,Y)$

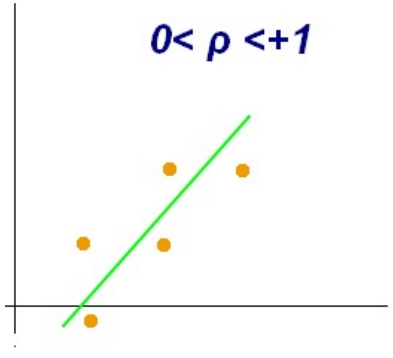
    - Second inequality
    - $0 \le E[(X^*-Y^*)^2]$

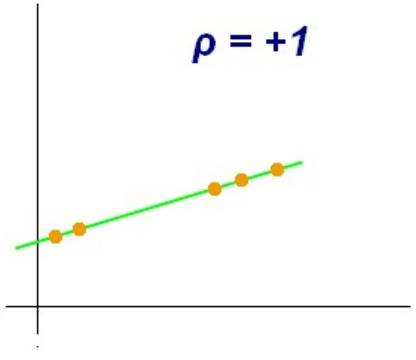
      $= E[(X^*)^2] + E[(Y^*)^2] - 2E[X^*Y^*]$

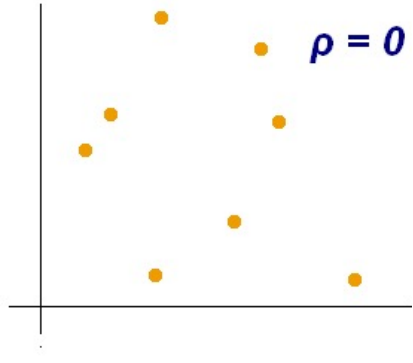
      $= 2(1 - Cor(X,Y))$
    - So, $Cor(X,Y) \le 1$



# Correlation

- **Property**: If X and Y are linearly related, i.e., Y = mX + c, and are non-constant (i.e., SD(X)>0 and SD(Y)>0), then |Cor(X,Y)| = 1
  - Proof:
    - When Y = mX + c, then SD(Y) = |m| SD(X)
    - Cor(X,Y)

    = Cov(X,Y) / (SD(X) SD(Y))

    = mVar(X) / (SD(X) |m|SD(X))

    = $\pm 1$

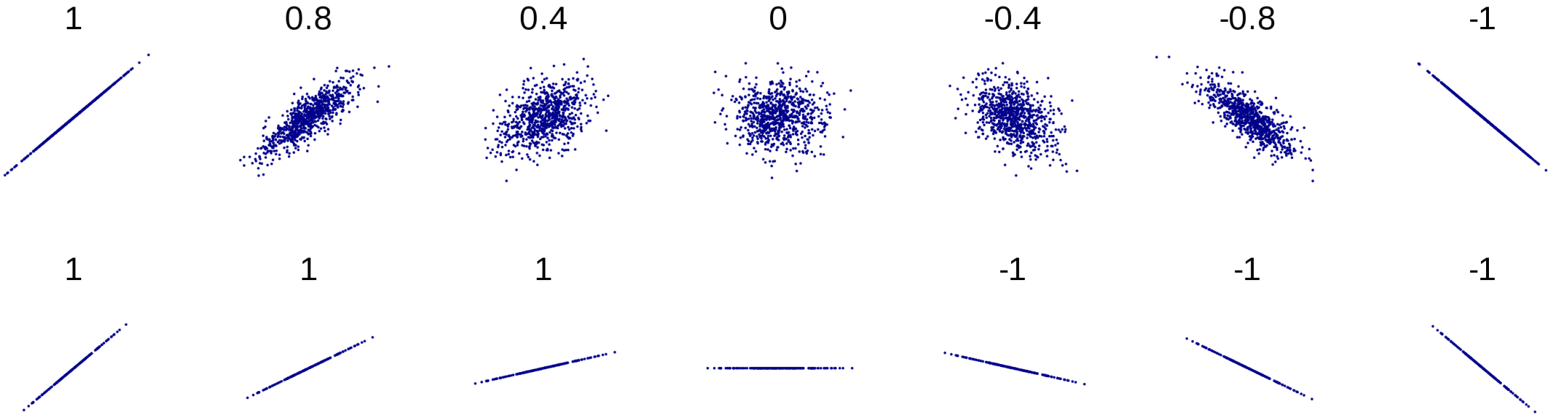
    = sign of the slope m

# Correlation

- **Property**: If $|Cor(X,Y)| = 1$, then X and Y are linearly related
  - Proof:
    - If $Cor(X,Y) = 1$, then $E[(X^*-Y^*)^2] = 2(1 - Cor(X,Y)) = 0$
      - For discrete X,Y: this must imply $X^*=Y^*$ for all $(x',y')$ where $P(X=x',Y=y') > 0$
        - Else the summation underlying the expectation cannot be zero
      - For continuous X,Y: this must imply $X^*=Y^*$ for all measures $(dx',dy')$ where $P(dx',dy') > 0$
        - $X^*$ and $Y^*$ can be unequal only on a countable set of isolated points where $P(dx',dy') > 0$
          - Else the integral underlying the expectation cannot be zero
    - If $Cor(X,Y) = (-1)$, then $E[(X^*+Y^*)^2] = 2(1 + Cor(X,Y)) = 0$
      - For discrete X,Y: this must imply $X^*=(-Y^*)$ for all $(x',y')$ where $P(X=x',Y=y') > 0$
      - For continuous X,Y: this must imply $X^*=(-Y^*)$ for all measures $(dx',dy')$ where $P(dx',dy') > 0$
        - Inequality can hold only on a countable set of isolated points where $P(dx',dy') > 0$
    - If $X^* = \pm Y^*$, then Y must be of the form $mX+c$

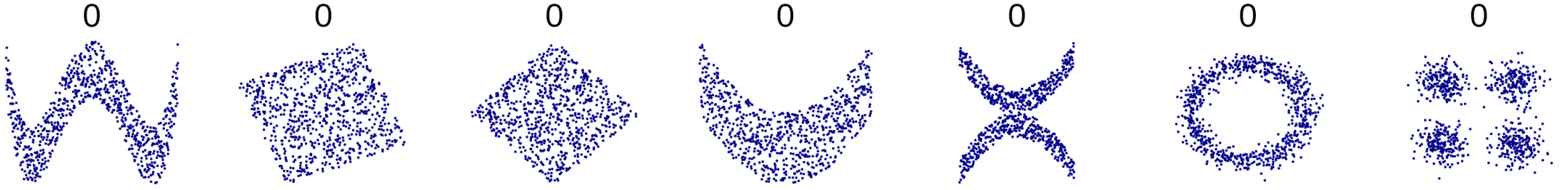
# Correlation

- Examples



# Correlation

- **Zero correlation doesn't imply independence**



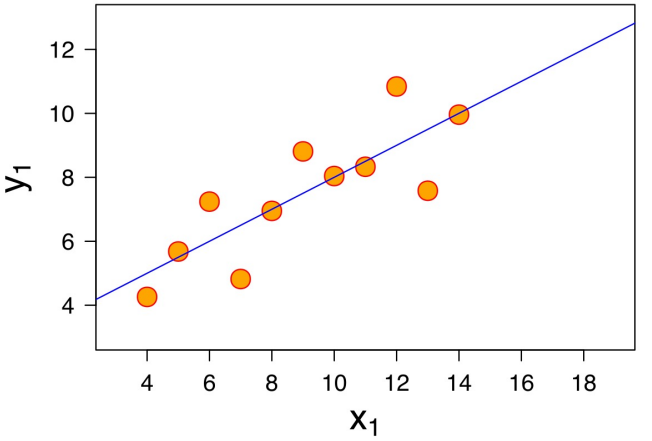

- Example: Let X be uniformly distributed within [-1,+1]. Let Y := $X^2$.
  - $Cov(X,X^2) = E[X.X^2] - E[X]E[X^2] = E[X^3] - 0.E[X^2] = 0$
  - Thus, Cov(X,Y) = 0 = Cor(X,Y) even though Y is a deterministic function of X

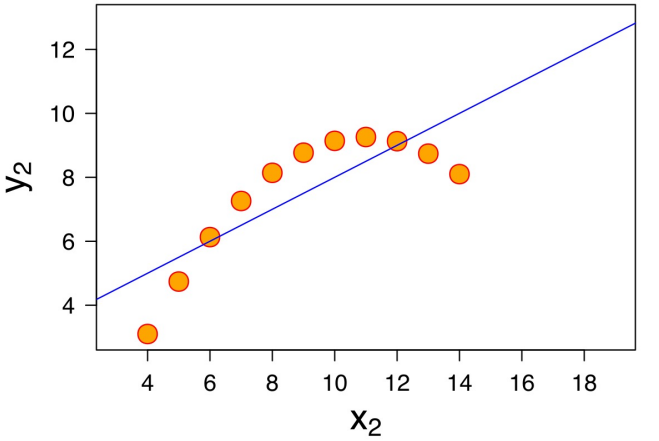
# Correlation

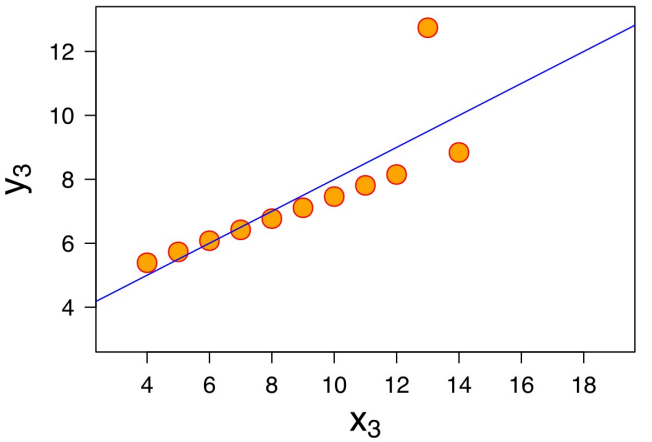
- If Y=mX+c, how to find the equation of line from data $\{(x_i,y_i): i=1,\ldots,n\}$?
  - Line must pass through (E[X],E[Y])
    - Because, when X=E[X], Y must be mE[X]+c, but that also equals E[Y]
  - We proved that: if Y=mX+c, then $|Cor(X,Y)|=1$ and $Y^* = \pm X^* = Cor(X,Y)\ X^*$
    - So, $(Y - E[Y]) / SD(Y) = Cor(X,Y)\ (X - E[X]) / SD(X)$
    - So, $Y = E[Y] + Cor(X,Y)\ (X - E[X])\ SD(Y) / SD(X)$
    - So, $Y = E[Y] + Cov(X,Y)\ (X - E[X]) / Var(X)$
    - This gives the equation of the line with:
      - Slope m := Cov(X,Y) / Var(X)
      - Intercept c := E[Y] – Cov(X,Y) E[X] / Var(X)

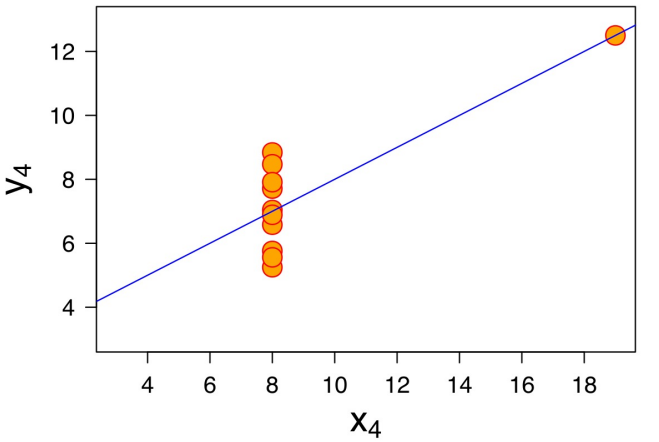
# Correlation

- Four sets of data with the same correlation of 0.816
  - Blue line is the line we'd have inferred if we assumed Y=mX+c



# Correlation

- **Non-zero correlation doesn't imply causation**
  - https://hbr.org/2015/06/beware-spurious-correlations
  - https://science.sciencemag.org/content/348/6238/980.2
  - http://www.tylervigen.com/spurious-correlations

BOOKS ET AL. | STATISTICS

## Spurious Correlations

+ See all authors and affiliations

*Science* 29 May 2015:
Vol. 348, Issue 6238, pp. 980
DOI: 10.1126/science.aac5518

Article | Figures & Data | Info & Metrics | eLetters | PDF

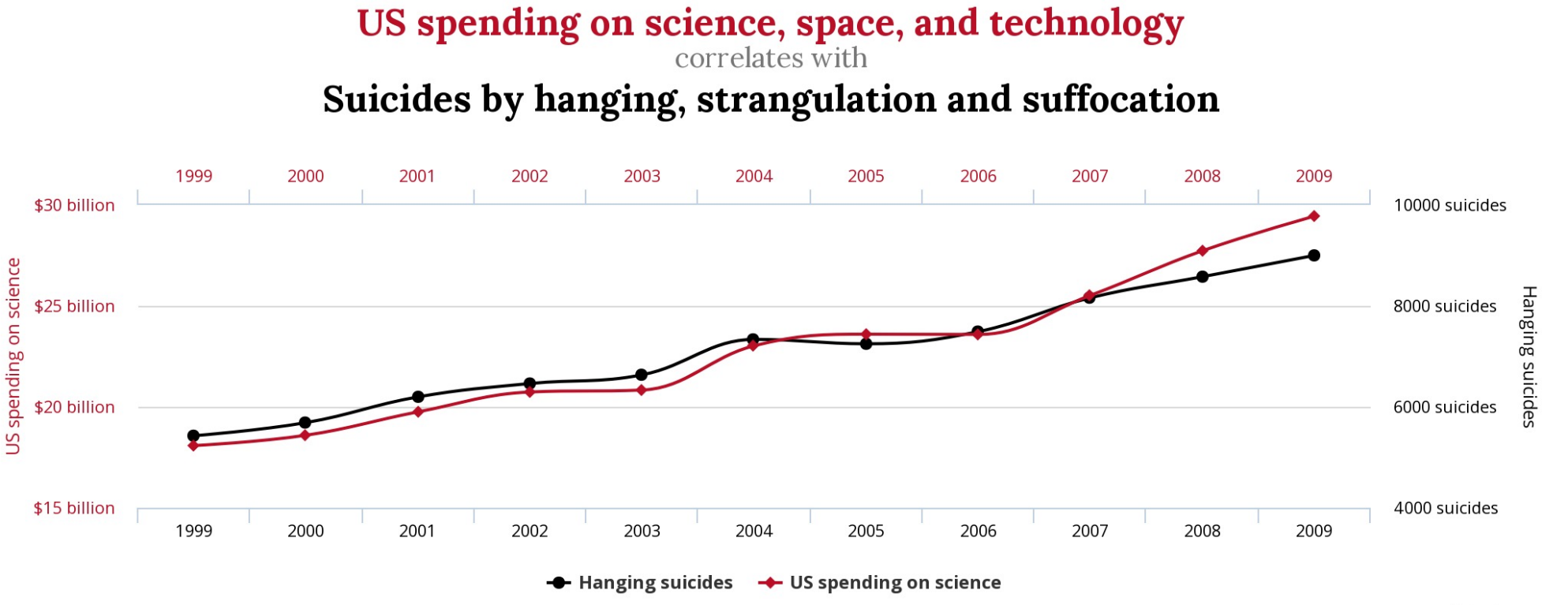
### Summary

Presented as a series of graphs prepared from real data sets, *Spurious Correlations* serves as a hilarious reminder that correlation most certainly does not equal causation.

View Full Text

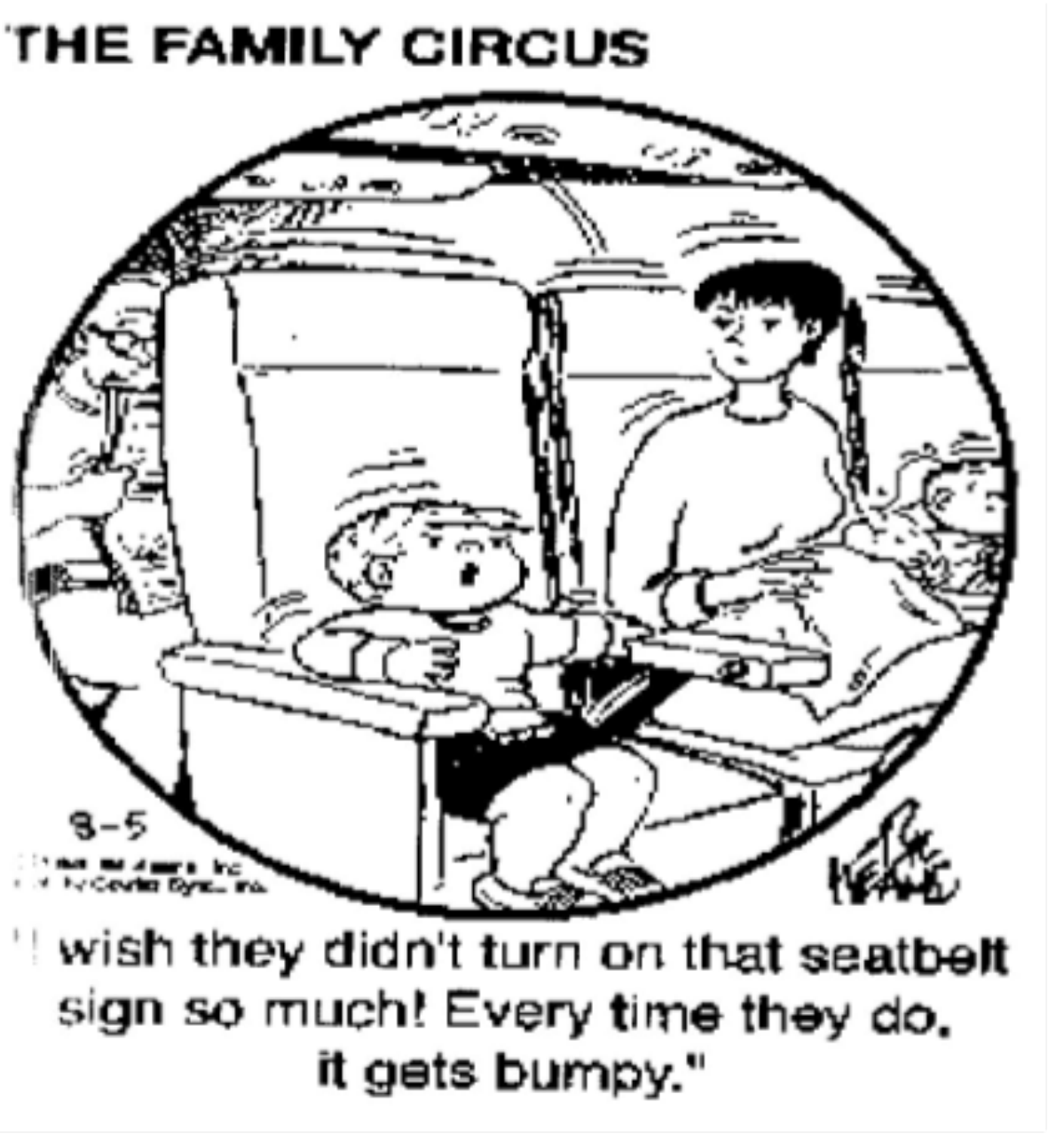
# Correlation

- Non-zero correlation doesn't imply causation
  - https://hbr.org/2015/06/beware-spurious-correlations
  - https://science.sciencemag.org/content/348/6238/980.2
  - http://www.tylervigen.com/spurious-correlations



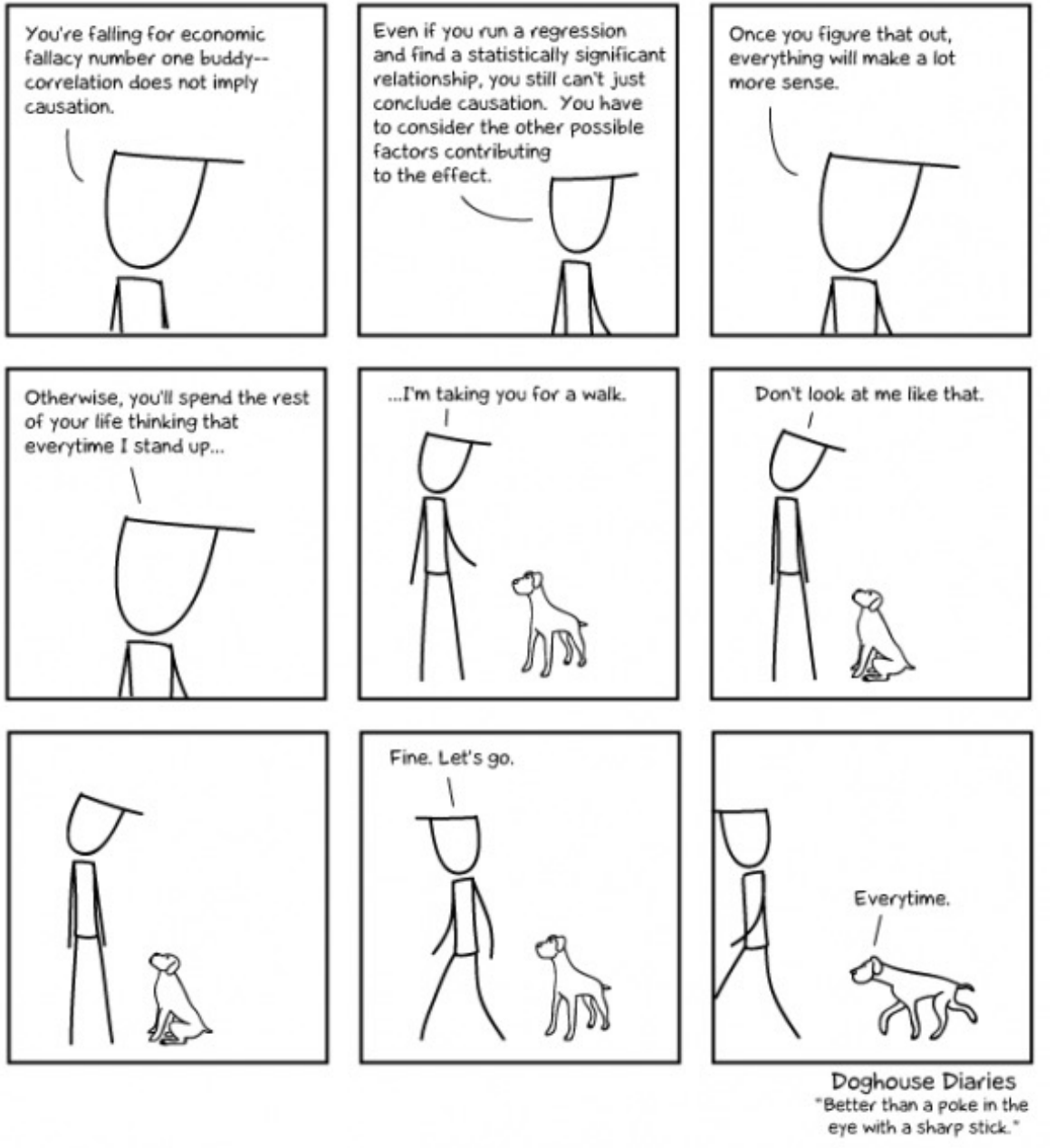
# Correlation

- Non-zero correlation doesn't imply causation



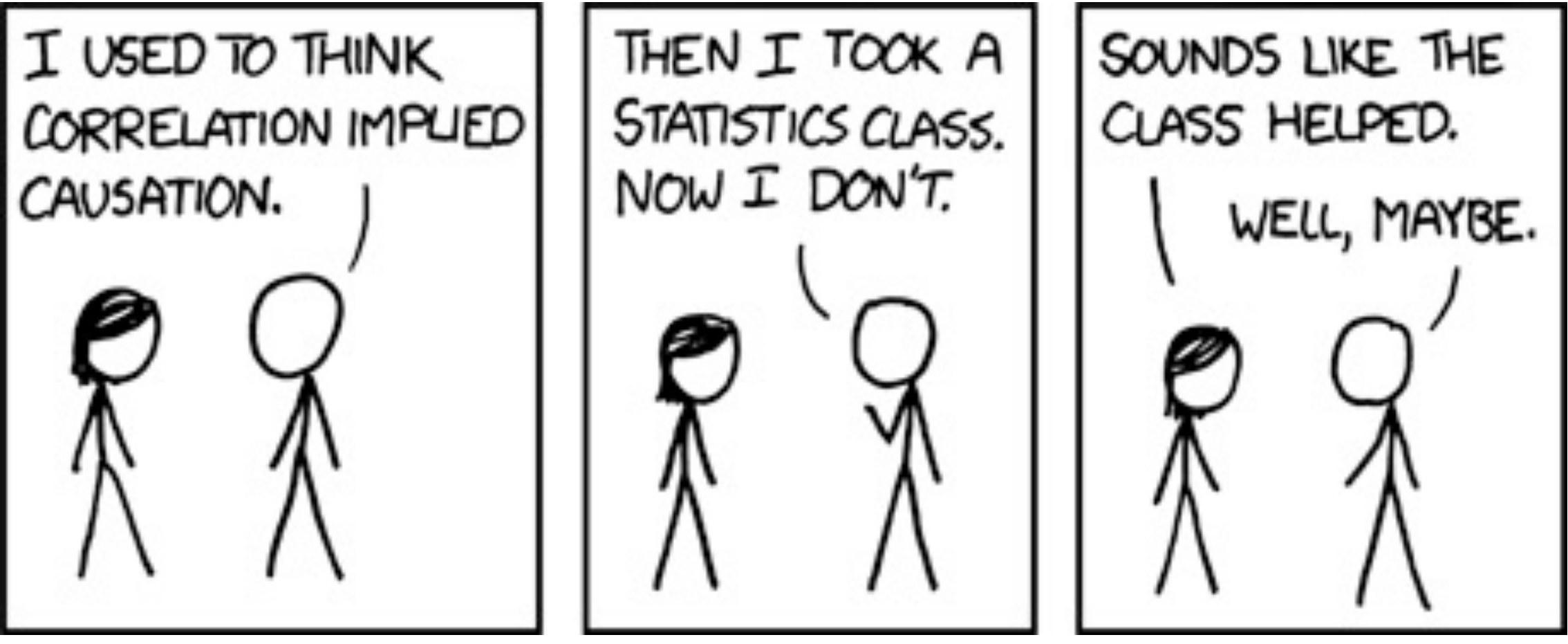
# Correlation

- Non-zero correlation doesn't imply causation



# Correlation

- Non-zero correlation doesn't imply causation



# Expectation in Life

- Action without expectation → Happiness [Indian Philosophy]