

Statistics



Statistics and data analysis

- ▶ **Analysis of data** is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.
- ▶ **Analysis of data** is a process of applying mathematical tools to extract useful and faithful information from observed data
- ▶ **Statistics** is a tool used in the process of analyzing data and in communicating and interpreting the results

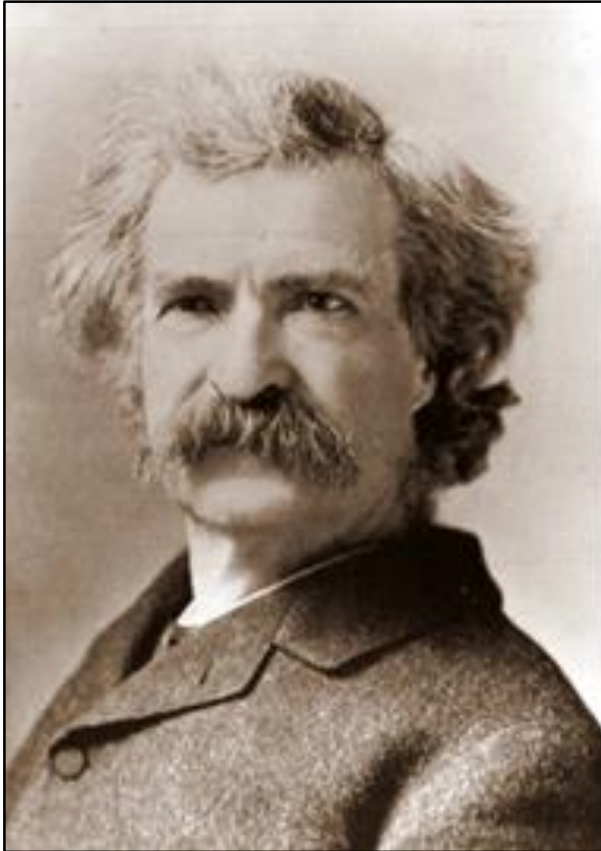
Statistics and data analysis in the age of computers

- ▶ The story of statistics is changing since:
 - ▶ More efficient algorithms and computers make deeper and more elaborate calculations possible and practical
 - ▶ The scope of data is changing in many respects, most notably volume

Statistics

- ▶ Statistics is a common bond supporting all other sciences as well as all quantitative social and business investigations
It provides standards of empirical proof and a language for communicating results in these domains

- ▶ The process of statistical investigation includes:
 - ▶ Designing experiments to maximize information
 - ▶ Using models to describe observations and assess their significance
 - ▶ Efficiently and effectively answering questions of interest
 - ▶ Verifying the validity of the process



**There are three kinds
of lies: lies, damned
lies, and statistics.
—Mark Twain
Chapters from My
Autobiography**

- ▶ The science of effectively drawing conclusions from data... OR...
The science of effectively lying

Statistics

- ▶ Can we make statistical arguments simple and convincing?
- ▶ Visualization and presentation
- ▶ Clear and simple statements
- ▶ Rigorous and accurate methodologies

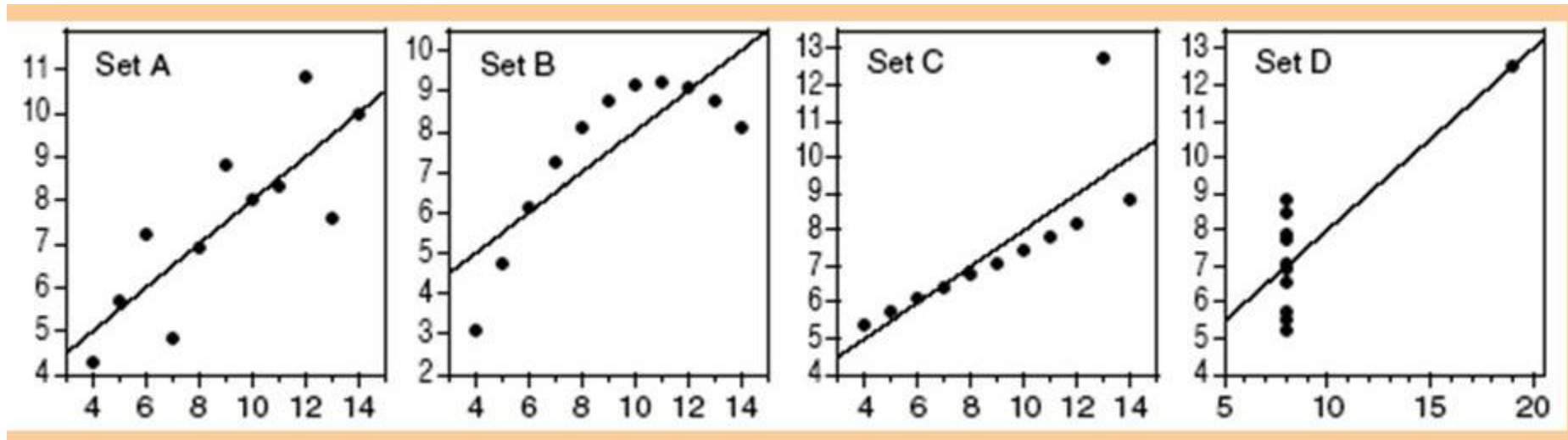


**"It's easier to
fool people than
to convince them
that they have
been fooled."**

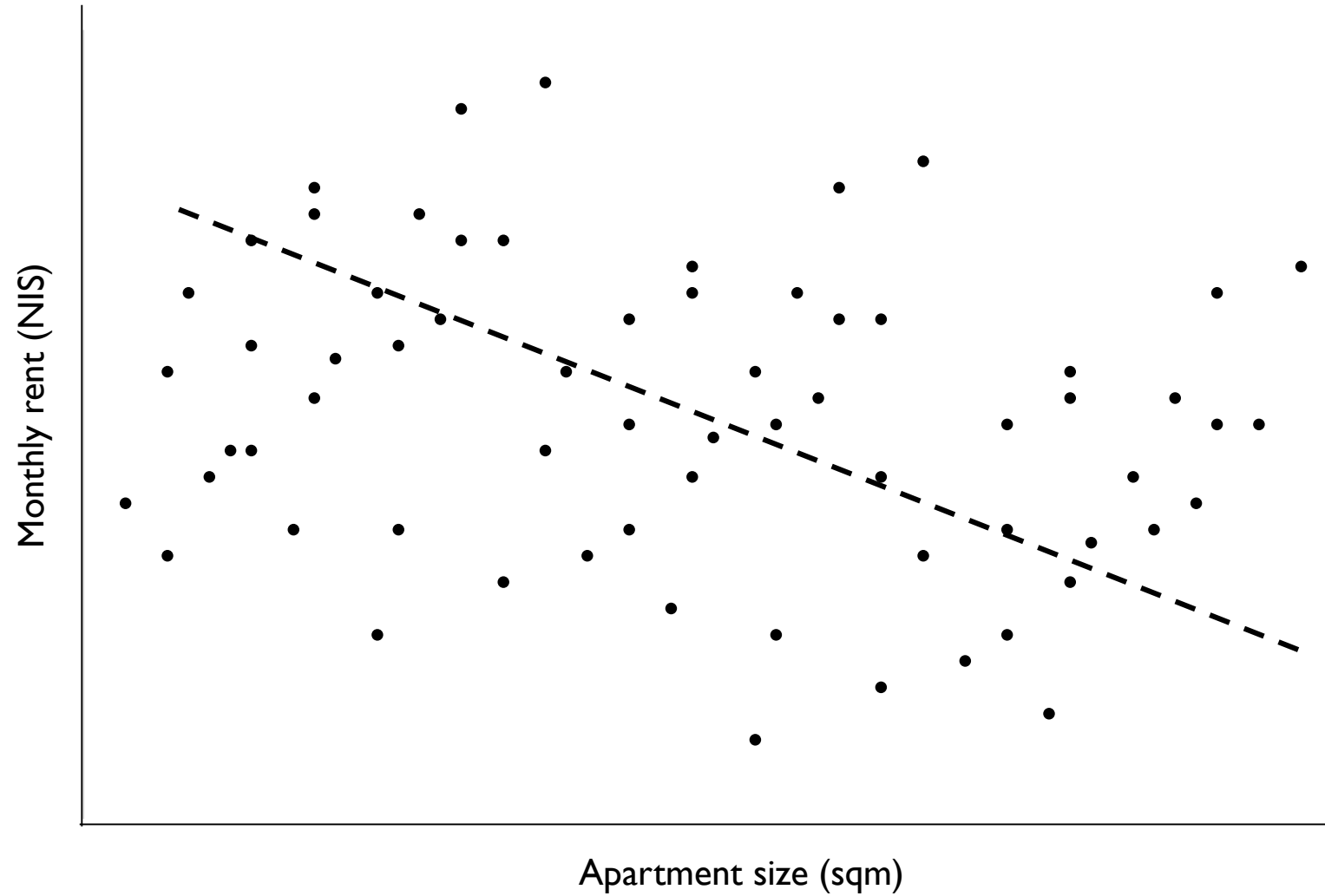
- Mark Twain -

Example 1

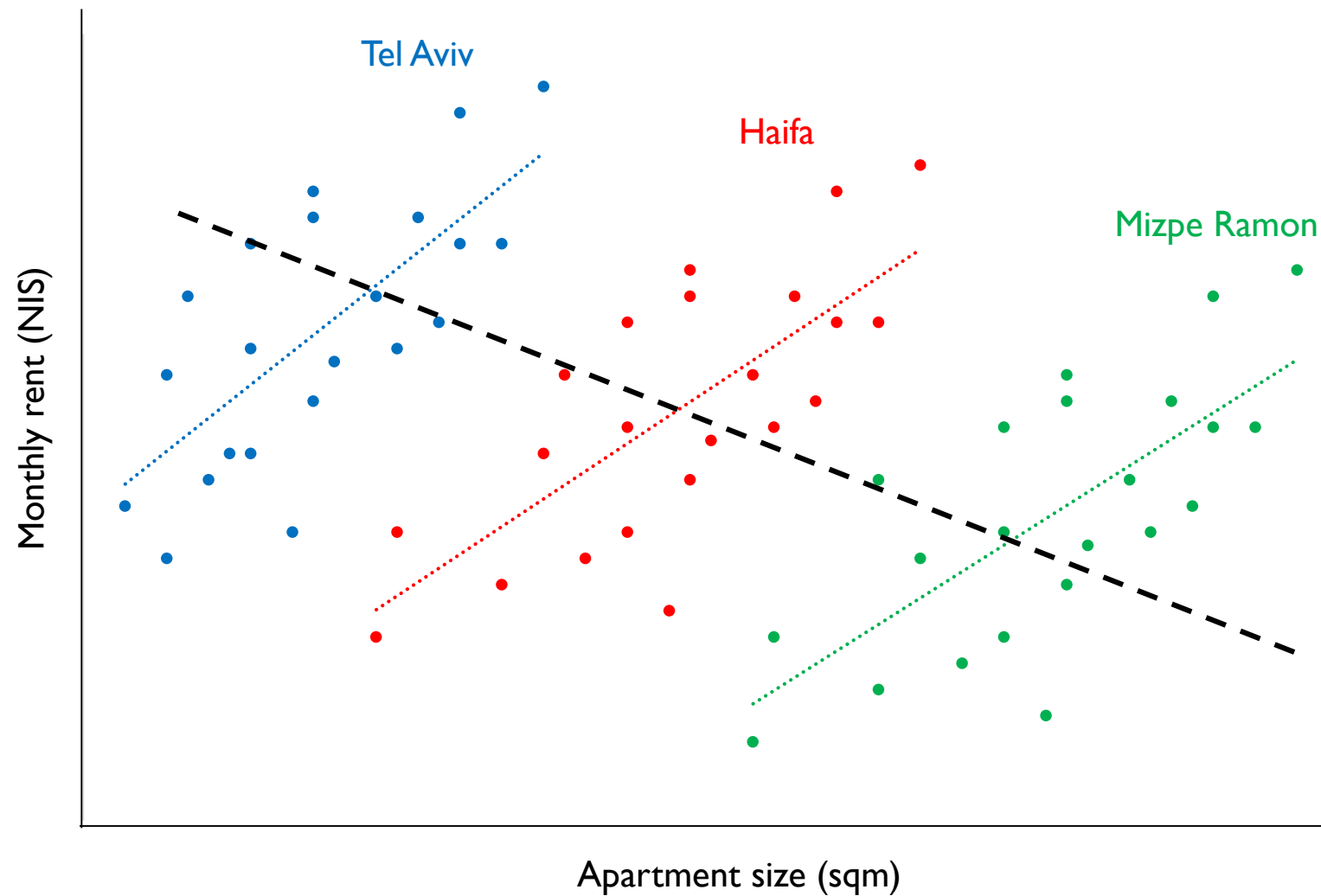
- ▶ Applying the same formal statistical tool on all of these yields the same result – high correlation between the variables
- ▶ But clearly – the actual conclusion should be different for each one of them



Example 2



Example 2



Example 3 – Simpson's Paradox

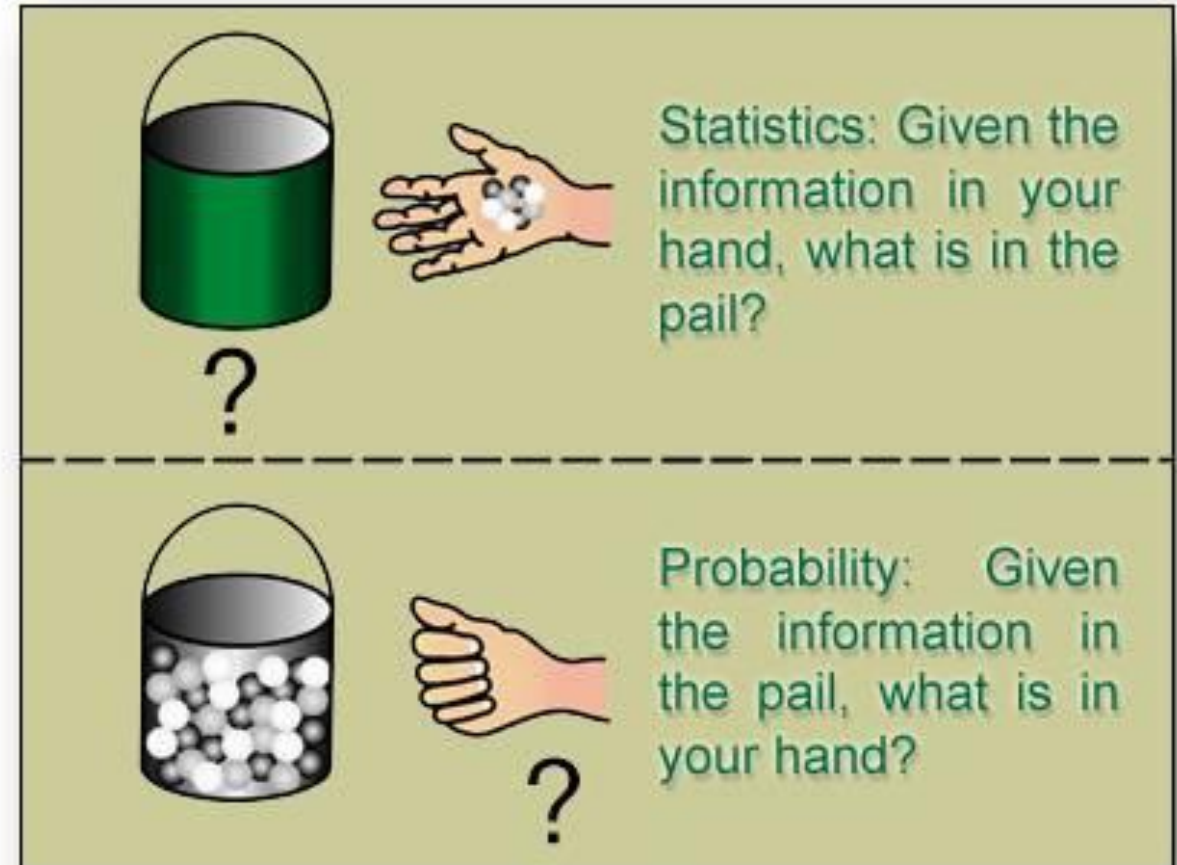
- ▶ The Randomistan basketball association conducted a test of its top leagues
- ▶ They recorded the number of basketball players who managed to score 5/5 free shots from the line
- ▶ The data is segmented by height and by gender

	Less than 1.70m	1.70-1.90	Taller than 1.90
Women	4/6	4/6	8/9
Men	1/2	1/2	23/27

- ▶ Who scores better from the line in Randomistan – men or women?
- ▶ Simpson's Paradox, is a phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined

Probability theory and statistics

- ▶ Statistics – given observations, what can we say about the underlying mechanism/system that gave rise to these observations?
- ▶ Probability – assuming a model - what is the expected behavior of observations from the model?



Random Variables

- ▶ A Random Variable (RV) is a numerical function defined on the probability sample space
- ▶ For each element of a sample space, the random variable takes on exactly one value
- ▶ Random Variables are usually denoted using upper case letters (X, Y)
- ▶ Individual outcomes for an RV are usually denoted using lower case letters (x, y)
- ▶ Example:
 - ▶ Toss a coin 5 times.
 - ▶ The sample space – Ω = all possible outcomes: 00000, 00001, 00010, 00011, ... , 01111, ..., 11101, 11110, 11111
 - ▶ One possible RV defined on this space – the outcome on the third toss
 - ▶ Another – count how many 1s (what is $P(Y=1)$ in this case?)
 - ▶ Is the number of 1s prime? (a binary RV)
 - ▶ Another – count how many 1s on even numbered tosses
 - ▶ Count how many 11 runs

Probability Distributions

- ▶ Probability Distribution: Table, Graph, or Formula that describes values a random variable can take on, and its corresponding probability (discrete RV) or density (continuous RV)
- ▶ Discrete Probability Distribution: Assigns probabilities (masses) to the individual outcomes
- ▶ Continuous Probability Distribution: Assigns density at individual points, probability of ranges can be obtained by integrating density function
- ▶ Discrete probability distribution: $p(y) = P(Y = y)$
- ▶ Continuous densities are denoted by $f(y)$.
We then have
$$P(Y \in I) = \int_I f(y) dy$$
- ▶ Cumulative Distribution Function: $F(y) = P(Y \leq y)$
- ▶ Probability distributions sum or integrate to 1
- ▶ What values can the CDF, F , of a random variable take?

Example – Rolling 2 Dice (Red/Green)

- ▶ Ω = All possible outcomes
- ▶ Y = Sum of the up faces of the two dice.
Table gives value of this function for all elements in Ω



Red\Green	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

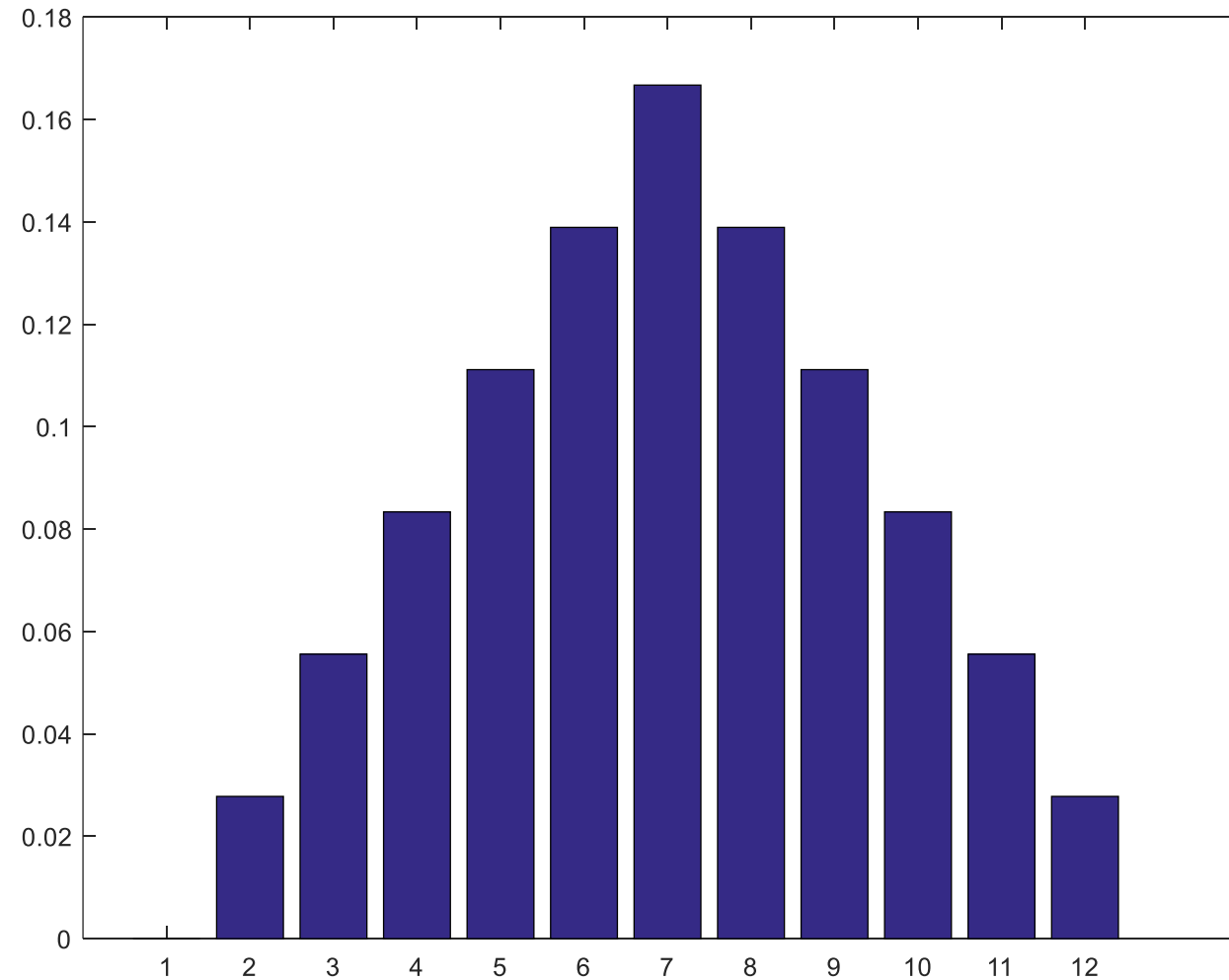
Rolling 2 Dice – Probability Mass Function & CDF

y	p(y)	F(y)
2	1/36	1/36
3	2/36	3/36
4	3/36	6/36
5	4/36	10/36
6	5/36	15/36
7	6/36	21/36
8	5/36	26/36
9	4/36	30/36
10	3/36	33/36
11	2/36	35/36
12	1/36	36/36

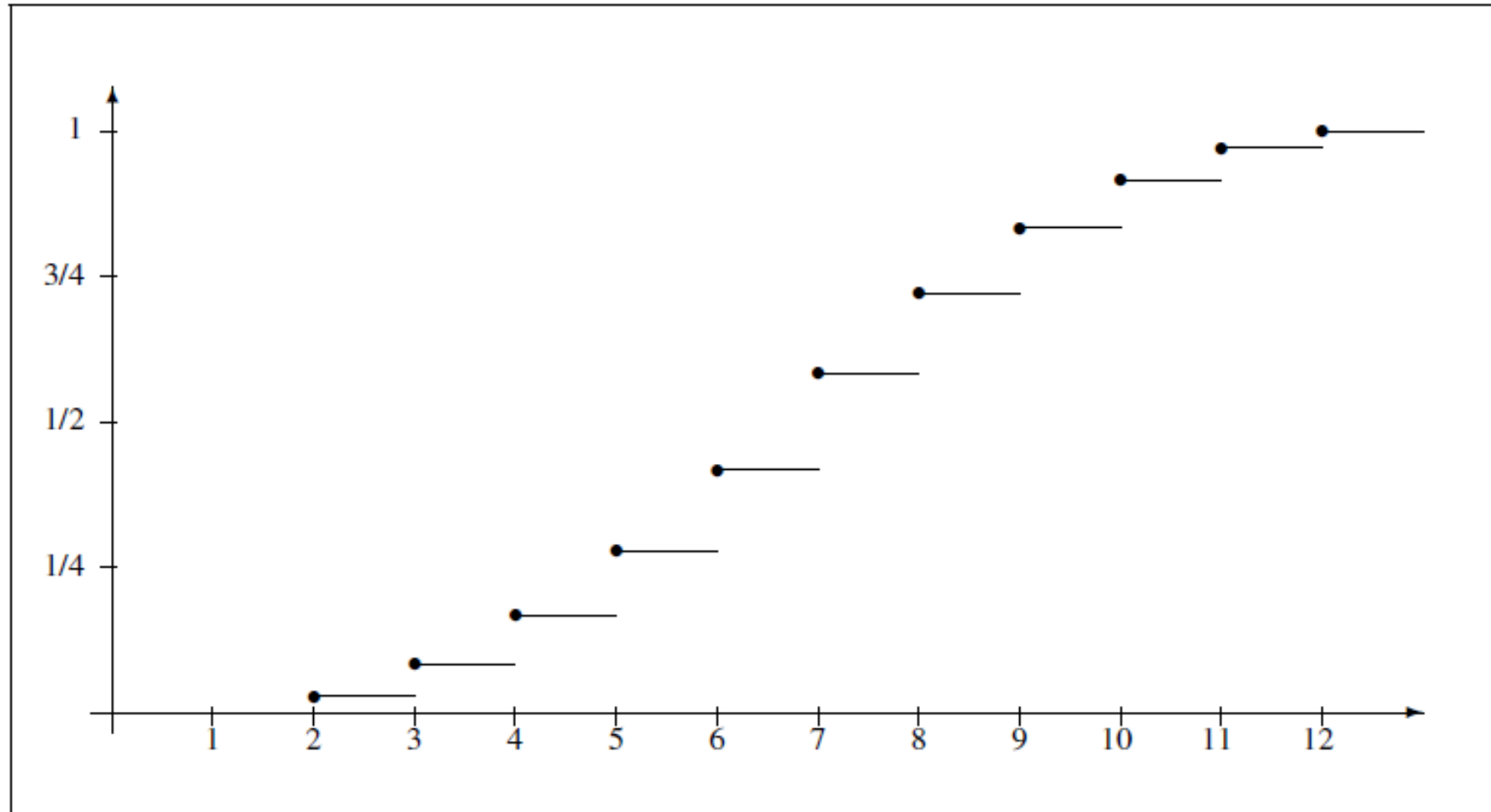
$$p(y) = \frac{\text{\# of ways 2 dice can sum to } y}{\text{\# of ways 2 dice configurations}}$$

$$F(y) = \sum_{t=2}^y p(t)$$

Rolling 2 Dice – Probability Mass Function



Rolling 2 Dice – Cumulative Distribution Function



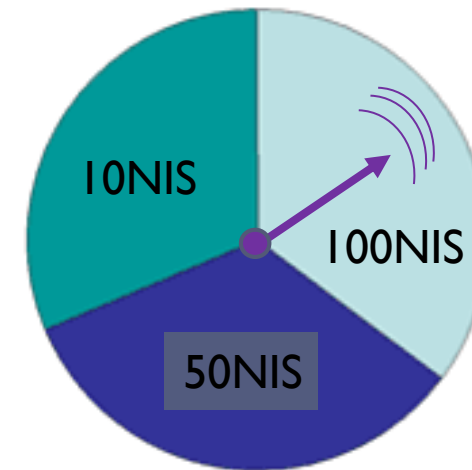
Exercise (1)

- ▶ Create a function that plot the PMF of the sum 3 dice
- ▶ Can you modify your function to plot the PMF for N dice where n is an input for the function?

Expected Values of Discrete RV's

- ▶ Mean (aka Expected Value) – the weighted average value on RV (or function of RV). Weighting is according to the underlying probability space
- ▶ Variance – Average squared deviation between a realization of an RV (or function of RV) and its mean
- ▶ Standard Deviation – Positive Square Root of Variance (in same units as the data)
- ▶ Notation:
 - ▶ Mean: $E(Y) = \mu$
 - ▶ Variance: $\text{Var}(Y) = \sigma^2$
 - ▶ Standard Deviation: σ

$$E(X) = \sum_{\text{all relevant } x} x p(x)$$



How much will we pay (or not) to play this game?

Expected Values of Discrete RV's

- ▶ Mean:

$$E(Y) = \mu = \sum_{all\ y} yp(y)$$

- ▶ Mean of a function g(Y):

$$E[g(Y)] = \sum_{all\ y} g(y)p(y)$$

- ▶ Variance:

$$\begin{aligned} V(Y) &= \sigma^2 = E[(Y - E(Y))^2] = E[(Y - \mu)^2] = \\ &= \sum_{all\ y} (y - \mu)^2 p(y) = \sum_{all\ y} (y^2 - 2\mu y + \mu^2) p(y) = \\ &= \sum_{all\ y} y^2 p(y) - 2\mu \sum_{all\ y} yp(y) + \mu^2 \sum_{all\ y} p(y) = \\ &= E[Y^2] - 2\mu(\mu) + \mu^2(1) = E[Y^2] - \mu^2 \end{aligned}$$

- ▶ Standard Deviation:

$$\sigma = +\sqrt{\sigma^2}$$

Expected Values of Linear Functions of Discrete RV's

- ▶ Linear Functions:

$$g(Y) = aY + b \quad (a, b \equiv \text{constants})$$

- ▶ Mean:

$$E[g(Y)] = E[aY + b] = \sum_{\text{all } y} (ay + b)p(y) = a \sum_{\text{all } y} yp(y) + b \sum_{\text{all } y} p(y) = a\mu + b$$

- ▶ Variance:

$$\begin{aligned} V(Y) &= V[aY + b] = \sum_{\text{all } y} ((ay + b) - (a\mu + b))^2 p(y) = \\ &= \sum_{\text{all } y} (ay - a\mu)^2 p(y) = \sum_{\text{all } y} [a^2(y - \mu)^2] p(y) = a^2 \sum_{\text{all } y} (y - \mu)^2 p(y) = a^2 \sigma^2 \end{aligned}$$

- ▶ Standard Deviation:

$$\sigma[aY + b] = |a|\sigma$$

Exercise (2)

- ▶ Create a function that rolling 2 dice 1000 times
- ▶ Store the sum of each 2 rolls
- ▶ Calculate the mean of all the 1000 sums
- ▶ Calculate the std of all the 1000 sums
- ▶ Are the results similar to what you expected?

Linearity of expectations

$$\begin{aligned} E(X + Y) &= \\ \sum_{x,y} p(x,y)(x + y) &= \\ \sum_{x,y} p(x,y)x + \sum_{x,y} p(x,y)y &= \\ \sum_x xp(x) + \sum_y yp(y) &= \\ E(X) + E(Y) \end{aligned}$$

Bernoulli Distribution

- ▶ An experiment consists of one trial
- ▶ It can result in one of 2 outcomes: Success or Failure (or a property being Present or Absent).
- ▶ Probability of Success ($Y = 1$) is p ($0 < p < 1$)
- ▶ Example: coin tossing

$$p(y) = \begin{cases} p & y = 1 \\ 1 - p & y = 0 \end{cases}$$
$$E(Y) = \sum_{y=0}^1 yp(y) = 0(1 - p) + 1p = p$$

$$E(Y^2) = 0^2(1 - p) + 1^2p = p$$

$$V(Y) = E(Y^2) - [E(Y)]^2 = p - p^2 = p(1 - p)$$

$$\sigma = \sqrt{p(1 - p)}$$

Binomial Experiment

- ▶ A binomial experiment consists of a series of n identical trials
- ▶ Each trial is Bernoulli as above
- ▶ Trials are independent (outcome of one has no bearing on outcomes of others)
- ▶ Probability of Success, p , is constant for all trials
- ▶ The random variable Y which counts the number of Successes in the n trials is said to follow Binomial Distribution with parameters n and p
- ▶ Y can take on the values $y=0,1,\dots,n$
- ▶ Notation: $Y \sim \text{Binom}(n,p)$

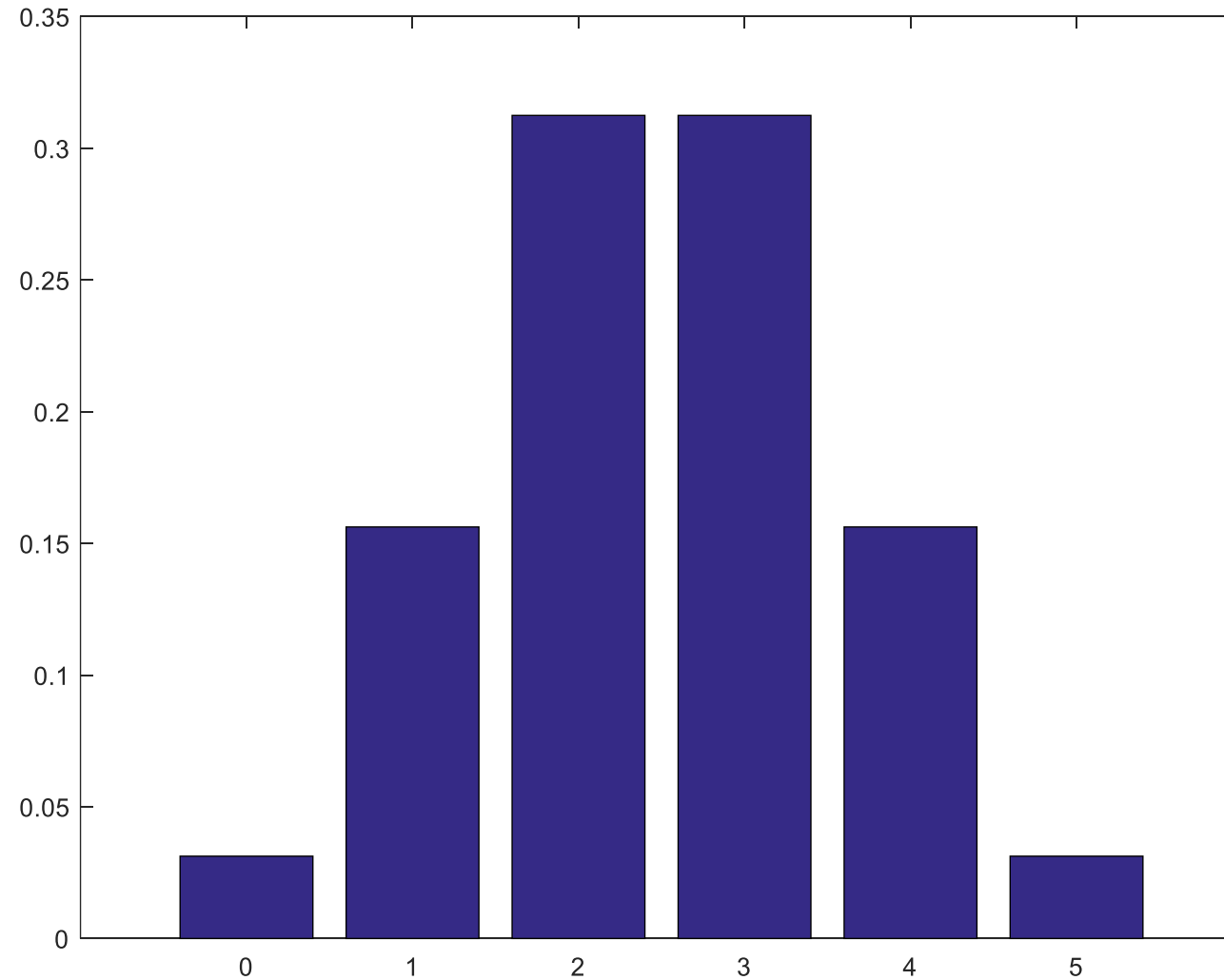
Binomial Distribution

$$P(Y = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

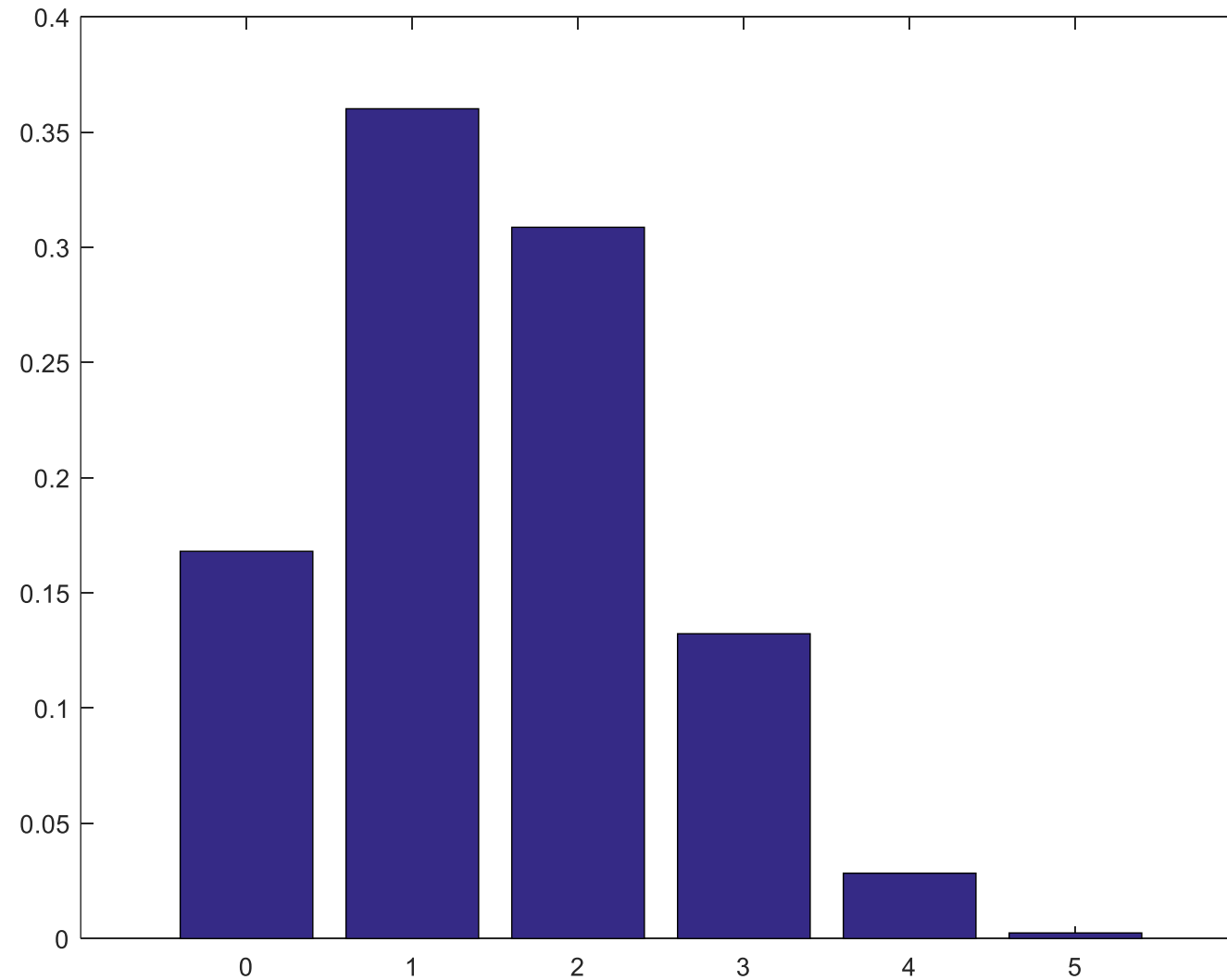
$$\sum_{k=0}^n P(Y = k) = \sum_{k=0}^n \binom{n}{k} p^k (1 - p)^{n-k} = (p + 1 - p)^n = 1$$



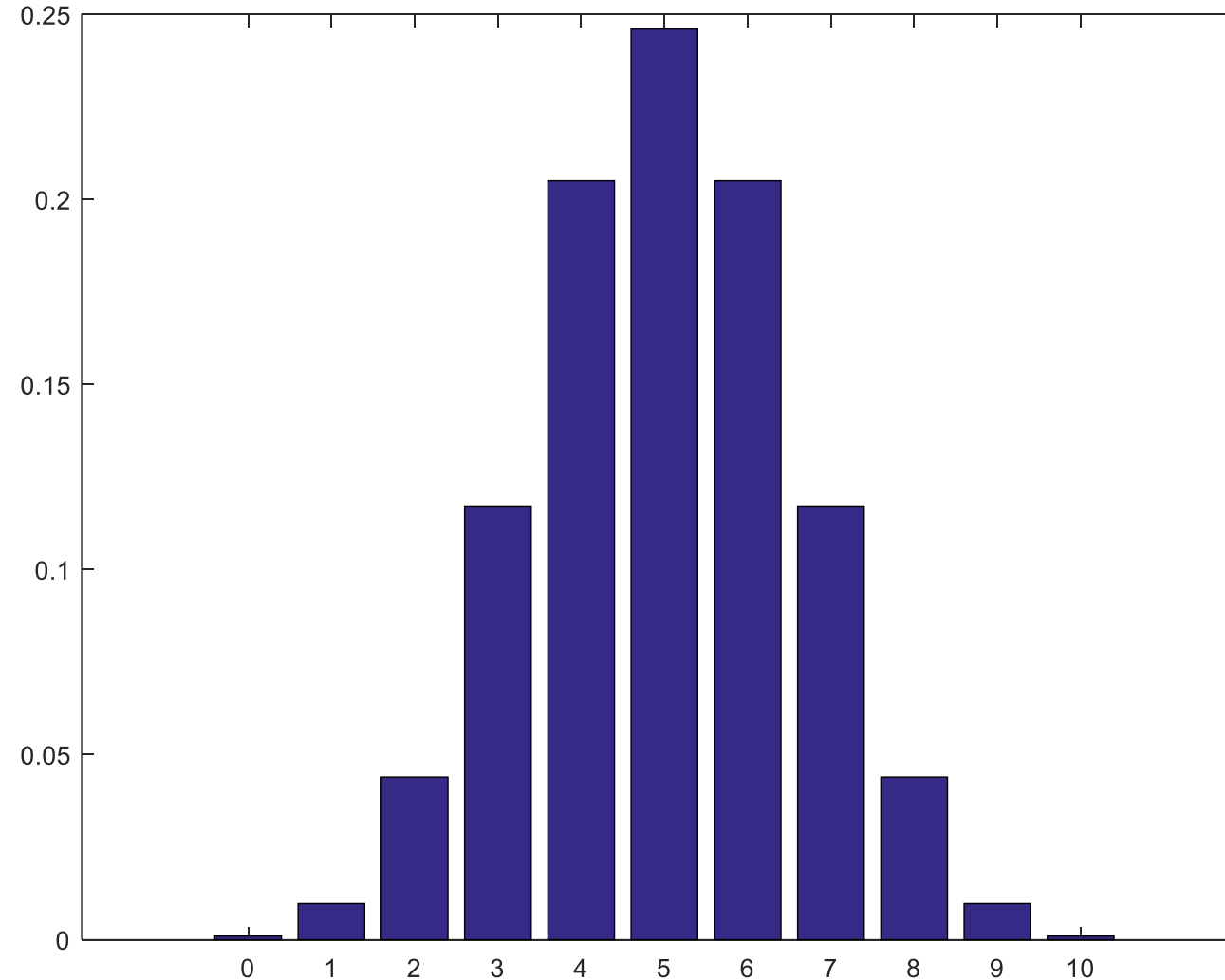
Binom(5,0.5) – tossing a fair coin 5 times, counting successes



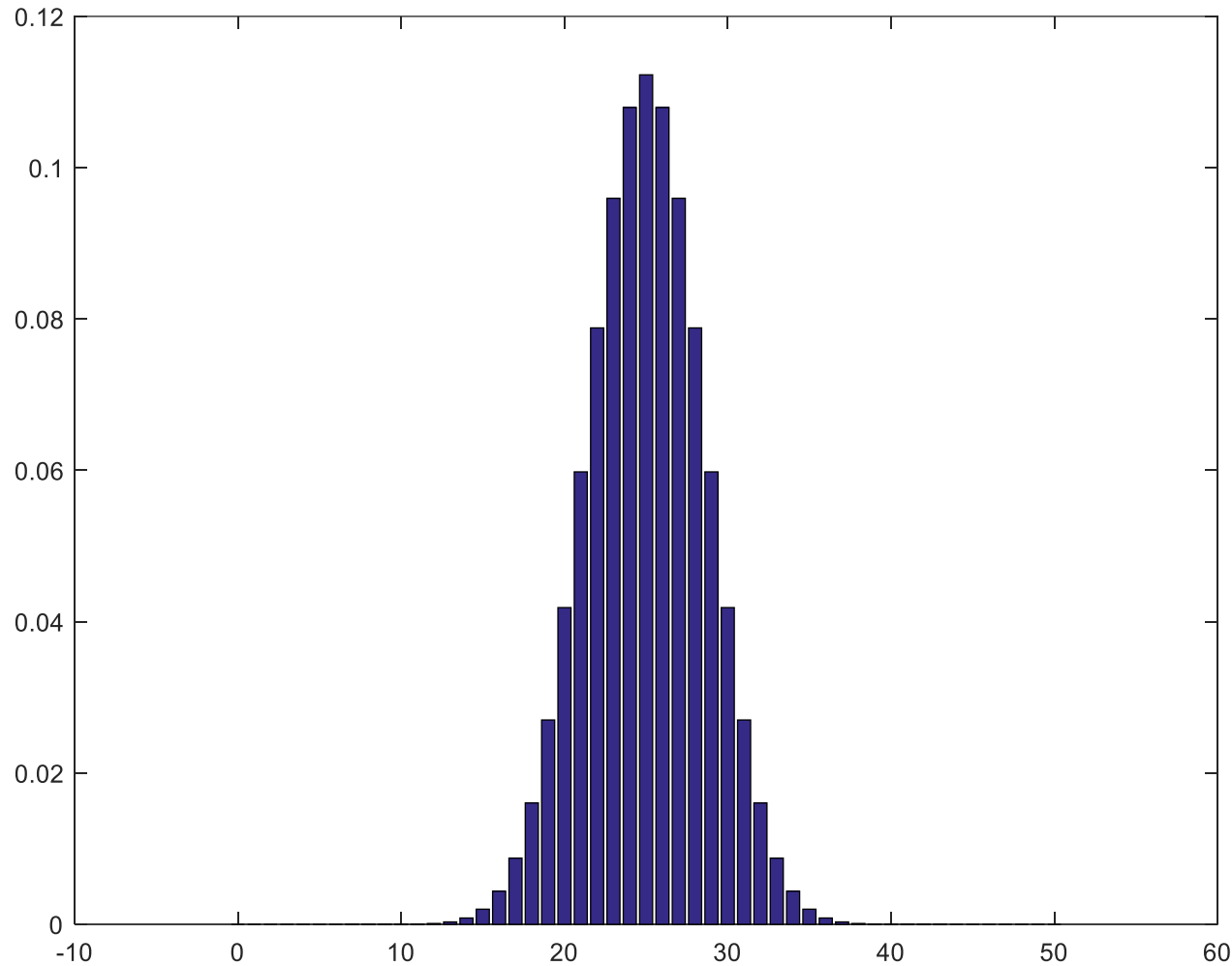
Binom(5,0.3)



Binom(10,0.5) – tossing a fair coin 10 times, counting successes

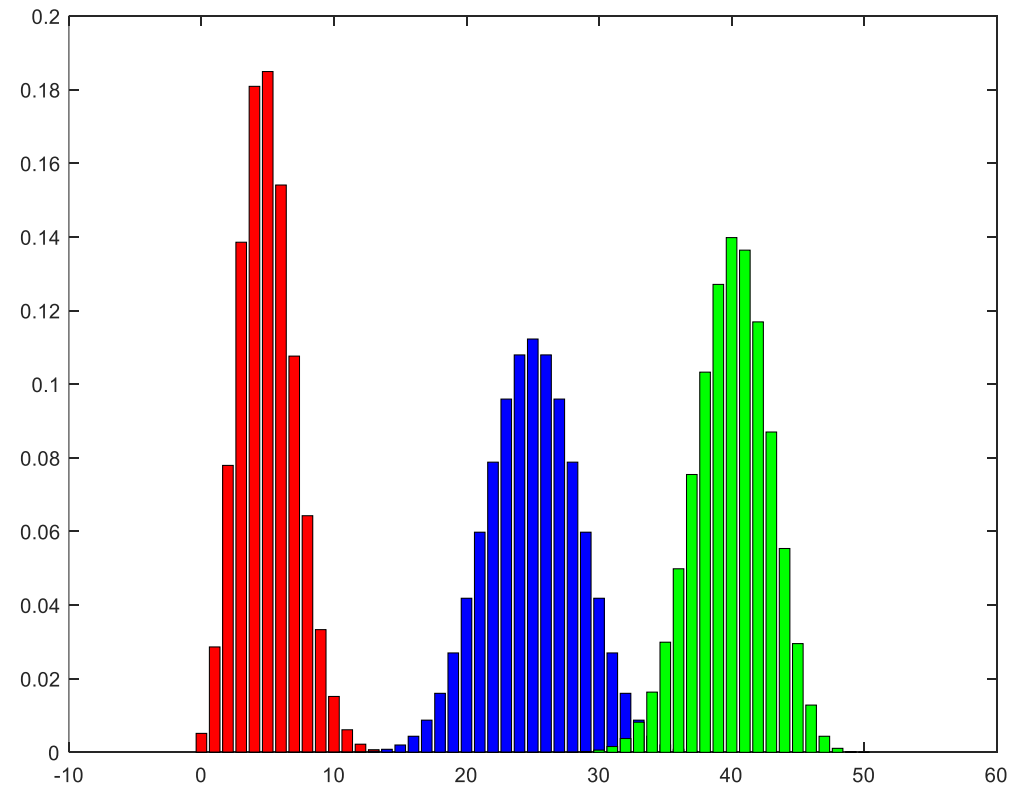


Binom(50,0.5) – tossing a fair coin 50 times, counting successes



Exercise (3)

- ▶ Create 3 different binomial distributions (different N and different p)
- ▶ Plot all of them on the same graph
- ▶ For example:



Binomial Distribution – Expected Value

$$f(y) = \frac{n!}{y!(n-y)!} p^y q^{n-y} \quad y = 0, 1, \dots, n \quad q = 1 - p$$

$$E(Y) = \sum_{y=0}^n y \left[\frac{n!}{y!(n-y)!} p^y q^{n-y} \right] = \sum_{y=1}^n y \left[\frac{n!}{y!(n-y)!} p^y q^{n-y} \right] =$$

$$= \sum_{y=1}^n \left[\frac{yn!}{y(y-1)!(n-y)!} p^y q^{n-y} \right] = \sum_{y=1}^n \left[\frac{n!}{(y-1)!(n-y)!} p^y q^{n-y} \right]$$

Let $y^* = y - 1 \rightarrow y^* + 1 = y$ Note: $y = 1, \dots, n \rightarrow y^* = 0, \dots, n - 1$

$$E(Y) = \sum_{y^*=0}^{n-1} \left[\frac{n(n-1)!}{y^*!(n-(y^*+1))!} p^{y^*+1} q^{n-(y^*+1)} \right] =$$

$$= np \sum_{y^*=0}^{n-1} \left[\frac{(n-1)!}{y^*!((n-1)-y^*)!} p^{y^*} q^{(n-1)-y^*} \right] =$$

$$= np(p+q)^{n-1} = np(p+(1-p))^{n-1} = np(1) = np$$

Binomial Distribution – Variance and S.D.

$$f(y) = \frac{n!}{y!(n-y)!} p^y q^{n-y} \quad y = 0, 1, \dots, n \quad q = 1 - p$$

Note: $E(Y^2)$ is difficult to get, but $E(Y(Y-1)) = E(Y^2) - E(Y)$ is not:

$$\begin{aligned} E(Y(Y-1)) &= \sum_{y=0}^n y(y-1) \left[\frac{n!}{y!(n-y)!} p^y q^{n-y} \right] = \sum_{y=2}^n y(y-1) \left[\frac{n!}{y!(n-y)!} p^y q^{n-y} \right] = \\ &= \sum_{y=2}^n \left[\frac{y(y-1)n!}{y(y-1)(y-2)!(n-y)!} p^y q^{n-y} \right] = \sum_{y=2}^n \left[\frac{n!}{(y-1)!(n-y)!} p^y q^{n-y} \right] \end{aligned}$$

Let $y^{**} = y - 2 \rightarrow y^{**} + 2 = y$ Note: $y = 2, \dots, n \rightarrow y^{**} = 0, \dots, n-2$

$$\begin{aligned} E(Y(Y-1)) &= \sum_{y^{**}=0}^{n-2} \left[\frac{n(n-1)(n-2)!}{y^{**}!(n-(y^{**}+2))!} p^{y^{**}+2} q^{n-(y^{**}+2)} \right] = \\ &= n(n-1)p^2 \sum_{y^{**}=0}^{n-2} \left[\frac{(n-2)!}{y^{**}!((n-2)-y^{**})!} p^{y^{**}} q^{(n-2)-y^{**}} \right] = \end{aligned}$$

$$= n(n-1)p^2(p+q)^{n-2} = n(n-1)p^2(p+(1-p))^{n-2} = n(n-1)p^2$$

$$E(Y^2) = E(Y(Y-1)) + E(Y) = n(n-1)p^2 + np = np((n-1)p + 1) = n^2p^2 - np^2 + np = n^2p^2 + np(1-p)$$

$$\begin{aligned} V(Y) &= E(Y^2) - [E(Y)]^2 = n^2p^2 + np(1-p) - n^2p^2 = np(1-p) \\ \sigma &= \sqrt{np(1-p)} \end{aligned}$$

Using the binomial distribution

Example: Experimental treatment for Kidney Cancer

- ▶ Suppose we have $n = 40$ patients who will be receiving an experimental therapy (Tx) which is believed to be better than current treatments (standard of care = SoC)
- ▶ The latter has a historically derived 5-year survival rate of 20%
- ▶ That is, under the SoC the probability of 5-year survival is $p = 0.2$
- ▶ We will now count 5-year survival under Tx and will then need to decide if the new experimental treatment is better.

Results and “The Question”

- ▶ Suppose that using the new treatment we find that 16 out of the 40 patients survive at least 5 years past diagnosis.
- ▶ Q: Does this result suggest that the new therapy, Tx, has a better 5-year survival rate than that of the SoC?

That is:

is the probability that a patient survives at least 5 years greater than 0.2 when treated using the new therapy?

What do we consider in answering the question of interest?

- ▶ We essentially ask ourselves the following:
 - ▶ If we assume that new therapy is **no better** than the current then what is the probability of seeing the observed numbers? That is – how likely are they to occur, in such case, by chance alone?
 - ▶ More specifically:
What is the probability of seeing 16 or more successes out of 40 if the success rate of the new therapy is also 0.2?
 - ▶ This is called estimating the p-value of the **OBSERVED RESULT** under the **NULL model**

Binomial...

- ▶ This is a binomial experiment situation...
 - ▶ There are $n = 40$ patients and we are counting the number of patients that survive 5 or more years
 - ▶ The individual patient outcomes are independent and under the NULL MODEL the probability of success is $p = 0.2$ for all patients
(that is: we assume that Tx is NOT better than the standard of care)

- ▶ So the random variable $X = \# \text{ of "successes" in the clinical trial}$ is, under the NULL model, Binomial with $n = 40$ and $p = 0.2$,
i.e. $X \sim \text{BIN}(40, 0.2)$

Example: Treatment of Kidney Cancer - cont

- ▶ $X \sim \text{BIN}(40, 0.2)$, find the probability that exactly 16 patients survive at least 5 years.

$$P(X = 16) = \binom{40}{16} * 20^{16} * 80^{24} = 0.001945$$

- ▶ This requires some calculator gymnastics and some scratchwork!
- ▶ But - keep in mind we need to find the probability of having **16 or more** patients surviving at least 5 yrs.

$$\sum_{x=16}^{40} P(X = x) = 0.002936$$

Conclusion (statistics helps decision making ...)

- ▶ Because it is highly unlikely ($p = 0.0029$) that we would see this many successes in a group of 40 patients if the new Tx had the same probability of success as the SoC we have to make a choice, either...
 - A. Tx's survival rate is less than 0.2 and we have obtained a very rare result by chance.
 - B. Our assumption about the success rate of the new Tx is wrong and in actuality it has a better than 20% 5-year survival rate making the observed result more plausible.
- ▶ Tx is better than the SoC treatment with p-value < 0.003 under a binomial null model
- ▶ The observed 5 yrs survival in Tx has a Right side p-value < 0.003 under a binomial null model that represents the SoC (Binom(40,0.2))

Exercise (4)

- ▶ In a manufacturing pipeline products are 1% defective. We are interested in examining a defective product to see what goes wrong on the belt. We need to ask the facility manager to send us a set of independent samples for examination.
- ▶ How many independent samples should we ask for in order to have a 75% probability of having at least one defective product in the batch sent?
- ▶ Answer the same question but where:
 - ▶ Products are 4% defective and we want a 95% probability of at least one defective product in the batch
 - ▶ Products are 20% defective and we want a 90% probability of at least 5 defective products in the batch
 - ▶ Products are 10% defective and we want a 90% probability of at least 10 defective products in the batch;
Can you comment on the difference between the answer here and that of C2?

Geometric Distribution

► Counts the number of Bernoulli trials needed until the first success (S) occurs

► If the probability of success is p then

- First Success on Trial 1: $S \dots$, $\Rightarrow P(Y = 1) = p$
- First Success on Trial 2: $FS \dots$, $\Rightarrow P(Y = 2) = (1-p)p$
- First Success on Trial k : $F\dots FS \dots$, $\Rightarrow P(Y = k) = (1-p)^{k-1} p = q^{k-1} p$

$$p(y) = (1 - p)^{y-1} p \quad y = 1, 2, \dots$$

$$\sum_{y=1}^{\infty} p(y) = \sum_{y=1}^{\infty} (1 - p)^{y-1} p = p \sum_{y=1}^{\infty} (1 - p)^{y-1}$$

Setting $y^*=y-1$ and noting that $y=1,2,\dots \rightarrow y^*=0,1,\dots$

$$\sum_{y=1}^{\infty} p(y) = p \sum_{y^*=0}^{\infty} (1 - p)^{y^*} = p \left[\frac{1}{1 - (1 - p)} \right] = \frac{p}{p} = 1$$

Geometric Distribution – Expectation and variance

$$\begin{aligned}
 E(Y) &= \sum_{y=1}^{\infty} y [q^{y-1} p] = p \sum_{y=1}^{\infty} \frac{dq^y}{dq} = p \frac{d}{dq} \sum_{y=1}^{\infty} q^y = p \frac{d}{dq} \left[q \sum_{y=1}^{\infty} q^{y-1} \right] = \\
 &= p \frac{d}{dq} \left[\frac{q}{1-q} \right] = p \left[\frac{(1-q)(1) - q(-1)}{(1-q)^2} \right] = \frac{p((1-q) + q)}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}
 \end{aligned}$$

$$\begin{aligned}
 E(Y(Y-1)) &= \sum_{y=1}^{\infty} y(y-1) [q^{y-1} p] = pq \sum_{y=1}^{\infty} \frac{d^2 q^y}{dq^2} = pq \frac{d^2}{dq^2} \sum_{y=1}^{\infty} q^y = pq \frac{d^2}{dq^2} \left[q \sum_{y=1}^{\infty} q^{y-1} \right] = \\
 &= pq \frac{d^2}{dq^2} \left[\frac{q}{1-q} \right] = pq \frac{d}{dq} \frac{1}{(1-q)^2} = pq (-2(1-q)^{-3}(-1)) = \frac{2pq}{(1-q)^3} = \frac{2pq}{p^3} = \frac{2q}{p^2}
 \end{aligned}$$

$$\Rightarrow E(Y^2) = E(Y(Y-1)) + E(Y) = \frac{2q}{p^2} + \frac{1}{p} = \frac{2(1-p) + p}{p^2} = \frac{2-p}{p^2}$$

$$\Rightarrow V(Y) = E(Y^2) - [E(Y)]^2 = \frac{2-p}{p^2} - \left[\frac{1}{p} \right]^2 = \frac{2-p-1}{p^2} = \frac{1-p}{p^2} = \frac{q}{p^2}$$

$$\Rightarrow \sigma = \sqrt{\frac{q}{p^2}}$$

Geometric distribution – an application example

- ▶ The click through rate (CTR) of a banner in a website is 0.03. We are inspecting a stream of independent entries to the site
- ▶ What is the probability that the first CT occurs at the 5th entry? At the 56th entry?
- ▶ Let G be a geometric random variable w $p = 0.03$. We are interested in $P(G = 5)$ and $P(G = 56)$
- ▶ Lets do the first one:
 $P(G = 5) = 0.97^4 * 0.03 \sim 0.026$
- ▶ What is the probability of seeing the first CT after at least 10 entries?
 $P(G \geq 10) = 0.97^9 \sim 0.76$
- ▶ Entries were observed in 30 consecutive hours.
In 7/30 1 hour periods a CT occurred on or before the 9th entry.
Does this make sense?
What if this were true in 16/30 periods? 23/30 periods?
- ▶ What is the probability of seeing the first CT between the 10th and 20th entries?
- ▶ In how many periods do we expect this to happen in our 30 hours experiment?
Can we assign a p-value to a deviation from this expected behavior?

Negative Binomial Distribution

- ▶ In successive Bernoulli(p) instances, what is the distribution of the number of trials needed until the r^{th} success.
(the Geometric Distribution is equivalent to $r=1$)
- ▶ For this number to equal y we should have exactly $r-1$ successes in first $y-1$ trials, followed by a success

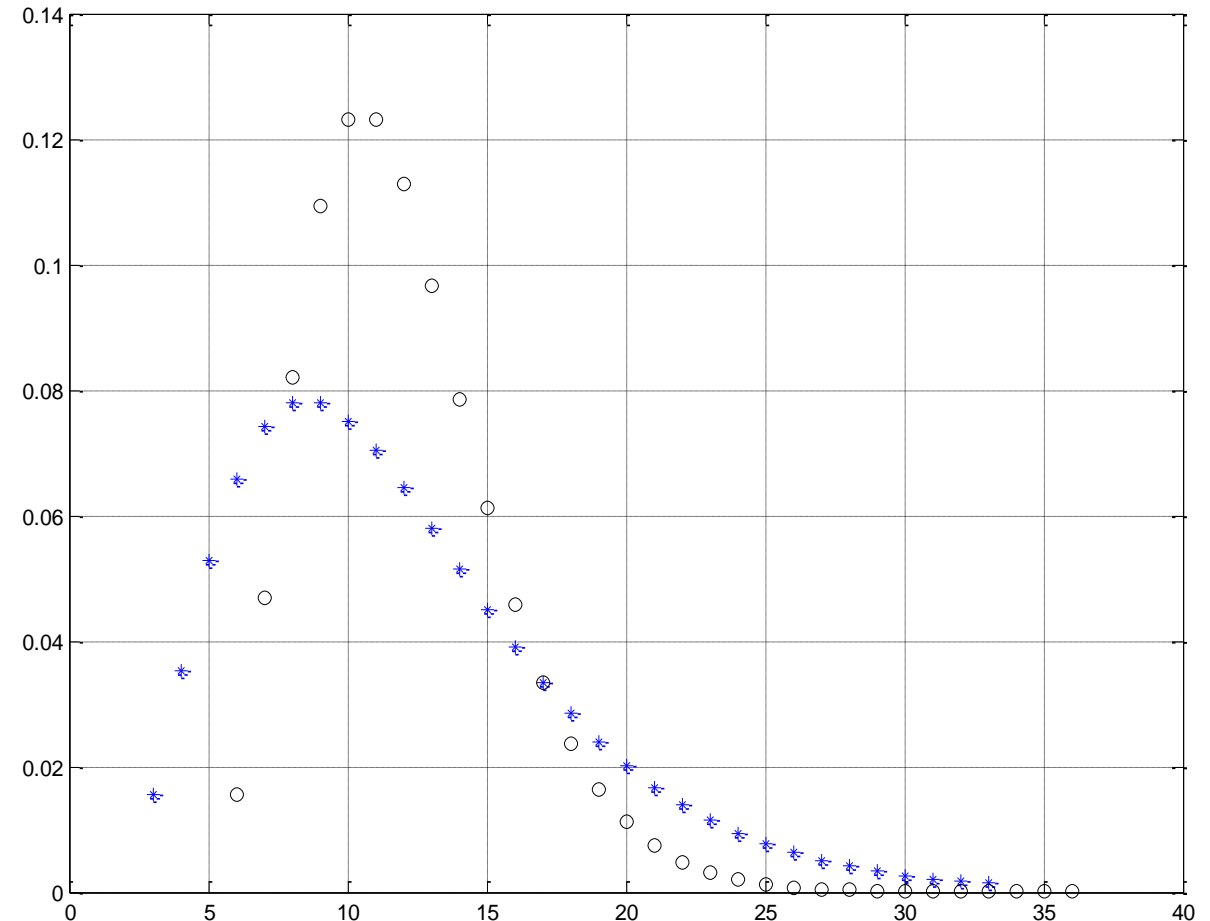
$$p(y) = \binom{y-1}{r-1} p^r (1-p)^{y-r} \quad y = r, r+1, \dots$$

$$E(Y) = \frac{r}{p} \quad (\text{Proof Given in Chapter 5})$$

$$V(Y) = \frac{r(1-p)}{p^2} \quad (\text{Proof Given in Chapter 5})$$

Example

- ▶ In the following figure we present the pdfs of two negative binomial distributions
One with $r_1 = 3$ and $p_1 = ??$
The other with $r_2 = 6$ and $p_2 = 0.5$
Determine which is which
- ▶ Given that the two distributions have the same mean, what is p_1 ?



Another Randomistan basketball story

- ▶ Players shoot synchronously
- ▶ Player A:
 - ▶ Probability of scoring = $p < 1/2$
 - ▶ Shoots until he has r successes
 - ▶ $N(A)$ is the attempt when that happened
- ▶ Player B:
 - ▶ Probability of scoring = mp for some integer $1 < m$ so that $mp < 1$
 - ▶ Shoots until she has mr successes.
 - ▶ $N(B)$ is the attempt when that happened
- ▶ Which is higher $E(N(A))$ or $E(N(B))$?
- ▶ Which is higher $V(N(A))$ or $V(N(B))$?

Exercise (5)

- ▶ Plot on the same graph the pdf of both players
- ▶ Where $p=0.25$, $m=3$ and $r=4$

Poisson Distribution

- ▶ Distribution often used to model the number of incidences in some characteristic unit of time or space:
 - ▶ Arrivals of customers to a store within one hour
 - ▶ Numbers of flaws in a roll of fabric of a given length
 - ▶ Number of visitors to a website in one minute
 - ▶ Number of calls to a service center in 10 mins

Poisson – a limit of binomials with an increasing n and a fixed mean

- ▶ Consider

$$X_1 \sim \text{Binom}(1, \lambda) \text{ and } X_2 \sim \text{Binom}\left(2, \frac{\lambda}{2}\right)$$

- ▶ Which is larger:

$$P(X_1 \geq 1) \text{ or } P(X_2 \geq 1)$$

- ▶ $P(X_1 \geq 1) = P(X_1 = 1) = \lambda$

- ▶ $P(X_2 \geq 1) = 1 - P(X_2 = 0) = 1 - \left(1 - \frac{\lambda}{2}\right)^2 = \lambda - \left(\frac{\lambda}{2}\right)^2 < \lambda$

Poisson – a limit of binomials with an increasing n and a fixed mean

$$\begin{aligned} X_n &\sim \text{Binom}\left(n, \frac{\lambda}{n}\right) \\ P(X_n = k) &= \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \\ &= 1 * \frac{\lambda^k}{k!} * e^{-\lambda} * 1 \end{aligned}$$

As $n \rightarrow \infty$

Poisson – a limit of binomials with an increasing n and a fixed mean

► So,

$$\forall k = 0, 1, \dots$$

► We have

$$P(X_n = k) \xrightarrow{n \rightarrow \infty} e^{-\lambda} \frac{\lambda^k}{k!}$$

Poisson Distribution

- ▶ $X \sim \text{Poisson}(\lambda)$, if

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

- ▶ Example – a website receives visits distributed as Poisson(0.5) per second
- ▶ What is the probability of no visits at a certain second?
- ▶ Answer: $\exp(-0.5)$
- ▶ What is the probability of no visits in a stretch of 10 seconds?
- ▶ Compute in two ways:
 - ▶ Poisson(5) yields $\exp(-5)$
 - ▶ 10 independent as above yields $(\exp(-0.5))^{10} = \exp(-5)$

Poisson Distribution – Expectation and Variance

$$f(y) = \frac{e^{-\lambda} \lambda^y}{y!} \quad y = 0, 1, 2, \dots$$

$$E(Y) = \sum_{y=0}^{\infty} y \left[\frac{e^{-\lambda} \lambda^y}{y!} \right] = \sum_{y=1}^{\infty} y \left[\frac{e^{-\lambda} \lambda^y}{y!} \right] = \sum_{y=1}^{\infty} \frac{e^{-\lambda} \lambda^y}{(y-1)!} = \lambda e^{-\lambda} \sum_{y=1}^{\infty} \frac{\lambda^{y-1}}{(y-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

$$\begin{aligned} E(Y(Y-1)) &= \sum_{y=0}^{\infty} y(y-1) \left[\frac{e^{-\lambda} \lambda^y}{y!} \right] = \sum_{y=2}^{\infty} y(y-1) \left[\frac{e^{-\lambda} \lambda^y}{y!} \right] = \sum_{y=2}^{\infty} \frac{e^{-\lambda} \lambda^y}{(y-2)!} \\ &= \lambda^2 e^{-\lambda} \sum_{y=2}^{\infty} \frac{\lambda^{y-2}}{(y-2)!} = \lambda^2 e^{-\lambda} e^{\lambda} = \lambda^2 \end{aligned}$$

$$\Rightarrow E(Y^2) = E(Y(Y-1)) + E(Y) = \lambda^2 + \lambda$$

$$\Rightarrow V(Y) = E(Y^2) - [E(Y)]^2 = \lambda^2 + \lambda - [\lambda]^2 = \lambda$$

$$\Rightarrow \sigma = \sqrt{\lambda}$$

Staff at the delivery room

- ▶ A maternity hospital needs to take a decision about staffing night (midnight-8AM) shifts
- ▶ For this purpose they want to estimate how many births would be expected during the night
- ▶ The hospital had 3300 deliveries in the past year, so if these happened randomly around the clock 1100 deliveries would be expected during the night shift
- ▶ The average number of deliveries per night is $1100/365$, which is ~ 3
- ▶ We then make the assumption that the number of babies per night is Poisson(3)
- ▶ Therefore – for example,
 - ▶ $P(0 \text{ deliveries per night}) = 3^0 e^{-3} / 0! \sim 0.05$
 - ▶ and
 - ▶ $P(2 \text{ deliveries per night}) = 3^2 e^{-3} / 2! \sim 0.23$
- ▶ Over the course of one year, what is the greatest number of deliveries that are expected to happen in at least one night?
 - ▶ (find maximal k such that $P(k) \geq 1/365$. *Ans: 8, what happens at 9?*)
- ▶ On how many days in the year would 5 or more deliveries be expected?
 - ▶ (*Ans: 68*)
- ▶ Management's staffing decision is to have sufficient staff to reduce the expected number of nights at which staff will be called from home to under 10. How many teams will be present in every night shift?
 - ▶ (find minimal k such that $CDF(k) \geq 1 - 10/365 = 0.9635$. *Ans: 6*)

Exercise (6)

- ▶ Use SciPy.stats to create two populations
 - ▶ Poisson with $\lambda = 35$, size=150,000 and shifted 18 to the right
 - ▶ Poisson with $\lambda = 10$, size=100,000 and shifted 18 to the right
- ▶ Concatenate the two populations
- ▶ Sample 500 observations from the unite population
- ▶ What is the population mean?
- ▶ What is the sample mean?
- ▶ Plot the population distribution
- ▶ Plot the sample distribution

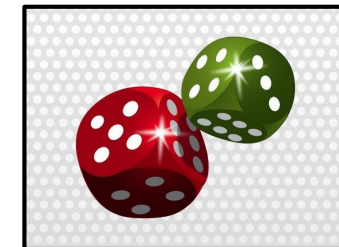
Statistical independence

► Example:

- Assuming that all outcomes have $P = 1/36$ is based on assuming that the result of one dice DOES NOT AFFECT the rolling of the other in any way
- What is the probability of $G = 3$ or 6 and $R = 5$?
- $P(G = 3 \text{ or } 6) = 1/3$
- $P(R = 5) = 1/6$
- The probability of the JOINT event is, according to the table on the left $1/36 + 1/36 = 1/18$
- This is just the product of the two probabilities:
 $P(G = 3 \text{ or } 6 \text{ and } R = 5) = 1/3 * 1/6 = 1/18$
- This is called STATISTICAL INDEPENDENCE
- When we defined $1/36$ in every entry we imply that the two rolls are independent random variables

Ω = All possible outcomes, that is:

(1,1) (1,2) (1,3) (1,4) (1,5) (1,6)
 (2,1) (2,2) (2,3) (2,4) (2,5) (2,6)
 (3,1) (3,2) (3,3) (3,4) (3,5) (3,6)
 (4,1) (4,2) (4,3) (4,4) (4,5) (4,6)
 (5,1) (5,2) (5,3) (5,4) (5,5) (5,6)
 (6,1) (6,2) (6,3) (6,4) (6,5) (6,6)



Definitions and factoids...

- ▶ Two events (subsets of the sample space Ω), A and B , are said to be statistically independent if the occurrence of one doesn't affect the occurrence of the other:
- ▶ $P(A | B) = P(A)$, where $P(A | B) = P(A \cap B) / P(B)$ is the conditional probability of A given B .
- ▶ From here we get
$$P(B | A) = P(A \cap B) / P(A) = P(A \cap B) P(B) / P(A) P(B) = P(A | B) P(B) / P(A) = P(B)$$
- ▶ Show that from here it follows that $P(A | B) = P(A | \neg B)$
- ▶ It also clearly follows that $P(A \cap B) = P(A)P(B)$

Independent random variables

- ▶ Two random variables X and Y , defined over the same space Ω have a joint distribution $p(x,y)$
- ▶ They also have marginal distributions
- ▶ The same marginal can often be joined (or coupled) in very different ways. The independent copula is only one of them
- ▶ They are called independent if for all numbers x and y we have $P(X = x \text{ and } Y = y) = P(X = x) * P(Y = y)$
- ▶ Or – for all x and y as above, the events $P(X = x)$ and $P(Y = y)$ are independent
- ▶ If X and Y are independent then $E(XY) = E(X)*E(Y)$
(prove this ...)
- ▶ Is the opposite true?

Linearity of expected values

- ▶ $E(X+Y) = E(X) + E(Y)$
- ▶ This is true for ANY random variables – they don't have to be independent
- ▶ This generalizes to any sums

- ▶ What about the Variance?

Covariance

- ▶ Consider X and Y defined on the same sample space Ω

$$\begin{aligned} \text{Cov}(X, Y) &= E\left((X - E(X))(Y - E(Y))\right) \\ &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

- ▶ Intuitively, the covariance between X and Y indicates how the values of X and Y move relative to each other
- ▶ If large values of X tend to happen with large values of Y , then $(X - E(X))(Y - E(Y))$ is positive on average
 - ▶ In this case, the covariance is positive and we say X and Y are positively correlated
- ▶ On the other hand, if X tends to be small when Y is large, then $(X - E(X))(Y - E(Y))$ is negative on average
 - ▶ In this case, the covariance is negative and we say X and Y are negatively correlated
- ▶ When X and Y are independent, what is $\text{Cov}(X, Y)$?
- ▶ Is the opposite true?

Variance

- ▶ What is the variance of sum of random variables?
- ▶ Consider X and Y defined on the same sample space Ω
- ▶ Let $Z=X+Y$

$$\begin{aligned}Var(Z) &= Cov(Z, Z) \\&= Cov(X + Y, X + Y) \\&= Cov(X, X) + Cov(X, Y) + Cov(Y, X) + Cov(Y, Y) \\&= Var(X) + Var(Y) + 2Cov(X, Y)\end{aligned}$$

- ▶ When X, Y are independent:
 $E(XY) = E(X)E(Y) \rightarrow Cov(X, Y) = 0 \rightarrow Var(X + Y) = Var(X) + Var(Y)$

Exercise (7)

- ▶ Define two random variables X and Y over the same probability space so that
 - ▶ $E(X) = 7$
 - ▶ $E(Y) = 63$
- ▶ And so that X and Y are:
 - ▶ not correlated ($\text{Cov}(X,Y) = 0$)
 - ▶ and NOT independent

Sums of independent random variables

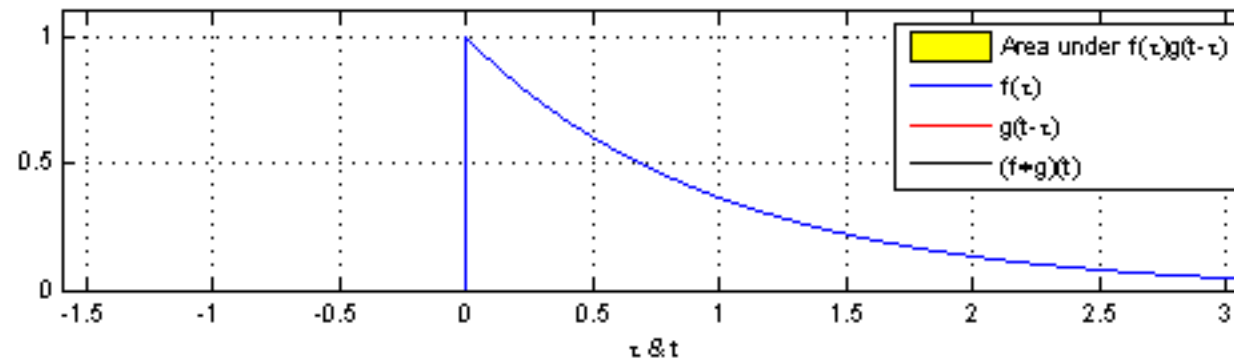
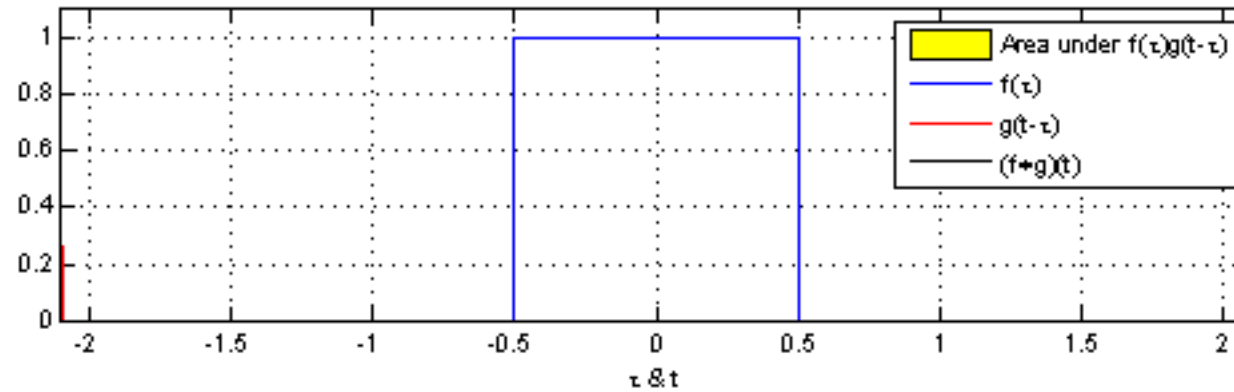
- ▶ Let X and Y be two independent random variables. Let $Z = X + Y$.
- ▶ Then

$$P(Z = z) = \sum_{i=-\infty}^{\infty} P(X = i)P(Y = z - i)$$

- ▶ For continuous random variables, the density function of Z is:

$$h(z) = \int_{-\infty}^{\infty} f(t)g(z - t)dt$$

Convolution



Exercise (8)

- ▶ Define two random variables X and Y that assume values on the non-negative integers so that:
 - ▶ Both X and Y assume at least two values with non-zero probability (they are not constant)
 - ▶ Let $Z = X+Y$. Then Z is uniformly distributed over the numbers $\{10, 11, 12, \dots, 101, 102\}$.

Sample/Coupon collection

- ▶ A website is seeking information about users from 10 different countries. It needs to observe the action of m users from each country to perform the analysis.
- ▶ How many visits will it take if every visit comes from each of the countries with equal probabilities and independent of all previous visits?
- ▶ At this point we will compute the expected value for the case $m=1$.
- ▶ We define random variables X_i , $i = 1 \dots 10$, as follows:
 - ▶ Let X_1 = the number of visits until the first country is in ($X_1 \geq 1$)
 - ▶ Let X_2 = the number of visits, after the first country is in, until the second country is in
 - ▶ ...
 - ▶ Let X_i = the number of visits, after the first $i-1$ countries are in, until the i -th country is also in.
- ▶ Now let
 - ▶ $T = X_1 + X_2 + X_3 + \dots + X_{10}$
- ▶ We are interested in
 - ▶ $E(T) = E(X_1) + E(X_2) + \dots + E(X_{10})$
- ▶ Since $X_i \sim \text{Geom}((10-i+1)/10)$ we can compute $E(X_i)$. It is $10/(10-i+1)$.
- ▶ So:
 - ▶ $E(T) = 10(1 + 1/2 + 1/3 + 1/4 + \dots + 1/10)$

Exercise (9)

- ▶ Let $n=3$ (3 countries) and $T=X_1+X_2+X_3$ (X_i the same as in section a), represent the time it takes to have seen all countries.
- ▶ Calculate the distribution of T in the range $1 \leq j \leq 6$. That is – compute $P(T=j)$ for all j s in the indicated range.

Sum of two independent Poissons is Poisson

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

Sum of 2 indpt Poisson

$$P(Y=k) = e^{-\mu} \frac{\mu^k}{k!}$$

X and Y indpt. Let $Z = X + Y$

$$P(Z=k) = \sum_{i=-\infty}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} \cdot e^{-\mu} \frac{\mu^{k-i}}{(k-i)!}$$

Here summands are 0 when either of the denominator factorials are negative

$$= e^{-(\lambda+\mu)} \cdot \frac{1}{k!} \sum_{i=0}^k \binom{k}{i} \lambda^i \mu^{k-i}$$

$$= e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^k}{k!}$$



The Hyper-Geometric Distribution (HG)

- ▶ In probability theory and statistics, the **hypergeometric distribution** is a discrete probability distribution that describes the probability of b successes (random draws for which the object drawn has a specified feature) in n draws, *without* replacement, from a finite population of size N that contains exactly B objects with that feature, wherein each draw is either a success or a failure

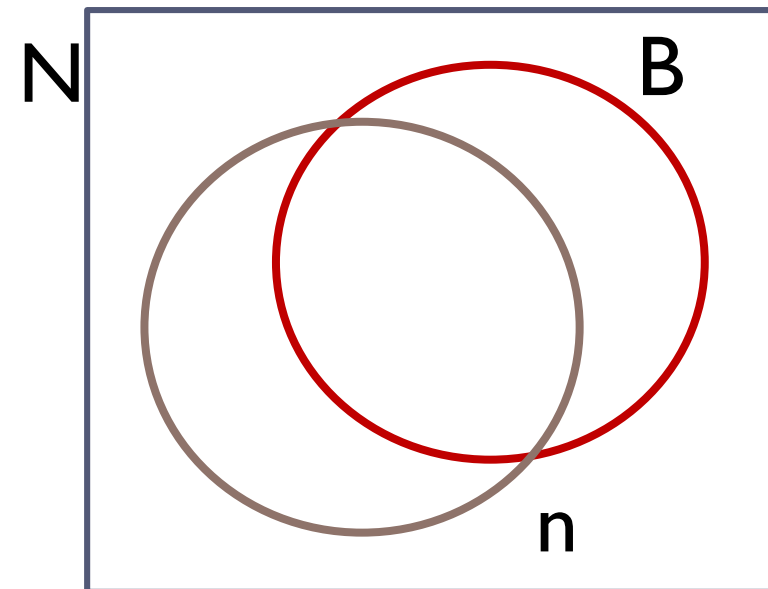
$$HG(N, B, n, b) = \frac{\binom{B}{b} \binom{N - B}{n - b}}{\binom{N}{n}}$$

- ▶ In contrast, the binomial distribution describes the probability of b successes in n draws *with replacement*

Example – effectiveness of a cold treatment

- ▶ $N = 200$
- ▶ $B = 105$ people whose cold lasted <3 days
- ▶ $n = 102$ people who took the Tx
- ▶ $b = 86$, the \cap
- ▶ We now compute the p-value for our observation under the HG null model – fix N and B and assume that all sets of size n are equiprobable:
- ▶ $HGT(N, B, n, b) = \sum_{t=b}^{\min(n, B)} HG(N, B, n, t)$
- ▶ And we conclude that the observed effectiveness of the drug has a p-value of $10^{(-21)}$.
- ▶ Quite effective

		Medicine Taken	
		yes	no
Cold Length	1 -3 days	86	19
	4 - 7 days	16	79
	Total	102	98
		Total	200



Winning the Lotto...

- ▶ 6 of 42 numbered balls are drawn at random without replacement
- ▶ You wrote 6 numbers on your lotto card ahead of time
- ▶ How many matches?
- ▶ The probability of no matches:

Your numbers

$$\frac{\binom{6}{0} \binom{36}{6}}{\binom{42}{6}} \approx 0.37$$

Lotto numbers

$$HG(42,6,6,0) = \frac{\binom{6}{0} \binom{36}{6}}{\binom{42}{6}}$$

Winning the Lotto...

- ▶ 6 of 42 numbered balls are drawn at random without replacement
- ▶ You wrote 6 numbers on your lotto card ahead of time
- ▶ How many matches?
- ▶ The probability of 1 match:

Your numbers

$$\frac{\binom{6}{1} \binom{36}{5}}{\binom{42}{6}} \approx 0.43$$

Lotto numbers

$$HG(42,6,6,1) = \frac{\binom{6}{1} \binom{36}{5}}{\binom{42}{6}}$$

Winning the Lotto...

- ▶ 6 of 42 numbered balls are drawn at random without replacement
- ▶ You wrote 6 numbers on your lotto card ahead of time
- ▶ How many matches?
- ▶ The probability of 5 matches:

Your numbers

$$\frac{\binom{6}{5} \binom{36}{1}}{\binom{42}{6}} \approx 0.000004$$

Lotto numbers

$$HG(42,6,6,5) = \frac{\binom{6}{5} \binom{36}{1}}{\binom{42}{6}}$$

Winning the Lotto...

- ▶ 6 of 42 numbered balls are drawn at random without replacement
- ▶ You wrote 6 numbers on your lotto card ahead of time
- ▶ How many matches?
- ▶ The probability of 6 match:

Your numbers

$$\frac{\binom{6}{6} \binom{36}{0}}{\binom{42}{6}} \approx 0.00000002$$

Lotto numbers

$$HG(42,6,6,6) = \frac{\binom{6}{6} \binom{36}{0}}{\binom{42}{6}}$$

Winning the Lotto...

- ▶ 6 of 42 numbered balls are drawn at random without replacement
- ▶ You wrote **7** numbers on your lotto card ahead of time
- ▶ How many matches?
- ▶ The probability of 6 match:

Your numbers

$$\frac{\binom{7}{6} \binom{36}{0}}{\binom{42}{6}} \approx 0.00000013$$

Lotto numbers

$$HG(42,7,6,6) = \frac{\binom{7}{6} \binom{36}{0}}{\binom{42}{6}}$$

Exercise (10)

- ▶ Lotto winning chances:
 - ▶ Let $X \sim \text{Hypergeometric}(N, B, n, *)$
 - ▶ One selects 6 numbers from 37 and 1 number from 7
 - ▶ He wins the big prize 10,000,000 if all 6 and the addition 1 matches the winning 6+1 numbers and smaller prize 500,000 if only the 6 the 6 out of 37 winning numbers
 - ▶ What is the probability of winning A?
 - ▶ of winning B?
 - ▶ What is the expected gain/loss of playing the Lotto if the cost of playing is 2.9?
 - ▶ And for 14 games?

Continuous probability distributions

- ▶ A continuous random variable can assume any value in an interval on the real line or in a collection of intervals
- ▶ It is not possible to talk about the probability of the random variable assuming a particular value
- ▶ Instead, we talk about the probability of the random variable assuming a value within a given interval
- ▶ The distribution is defined by a probability density function $p(x)$
- ▶ The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be the area under the graph of the probability density function between x_1 and x_2

Uniform Probability Distribution

A random variable is uniformly distributed whenever the probability is proportional to the length of the interval

- ▶ Uniform Probability Density Function

$$f(x) = 1/(b - a) \text{ for } a \leq x \leq b$$
$$= 0 \text{ elsewhere}$$

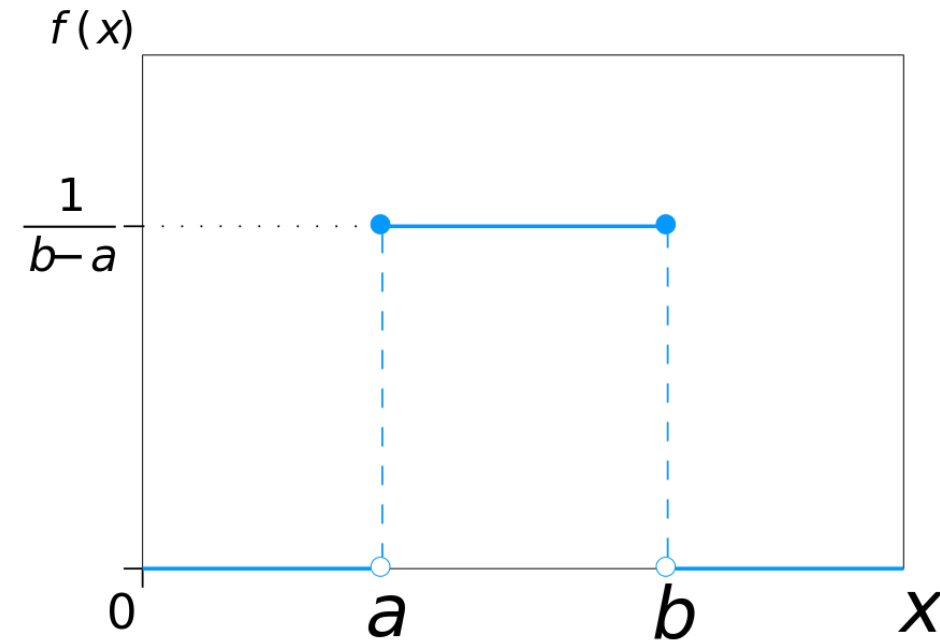
- ▶ Expected Value of x

$$E(x) = (a + b)/2$$

- ▶ Variance of x

$$\text{Var}(x) = (b - a)^2/12$$

where: a = smallest value the variable can assume
 b = largest value the variable can assume



CDF of continuous random variables

- ▶ The cumulative distribution function is

$$F(x) = \int_{-\infty}^x f(t)dt$$

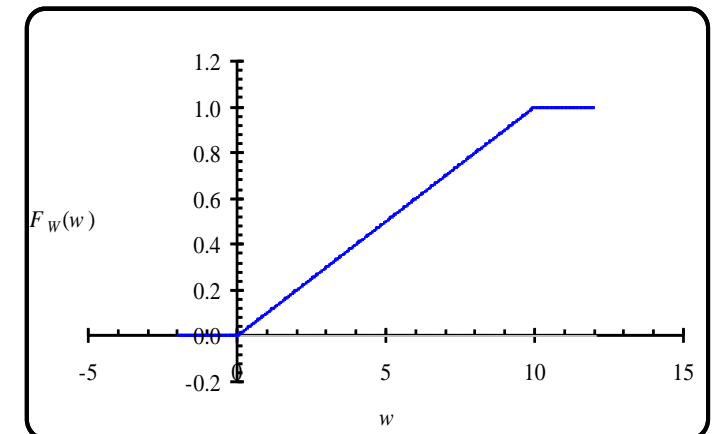
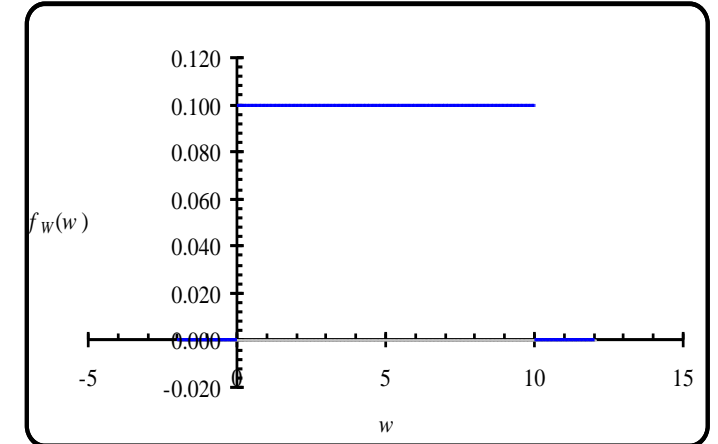
- ▶ and for any interval $I = [a,b]$ we can now have two expressions for

$P(I) = P(X \text{ assumes a value in the interval})$

$$P(I) = \int_a^b f(t)dt$$

- ▶ and

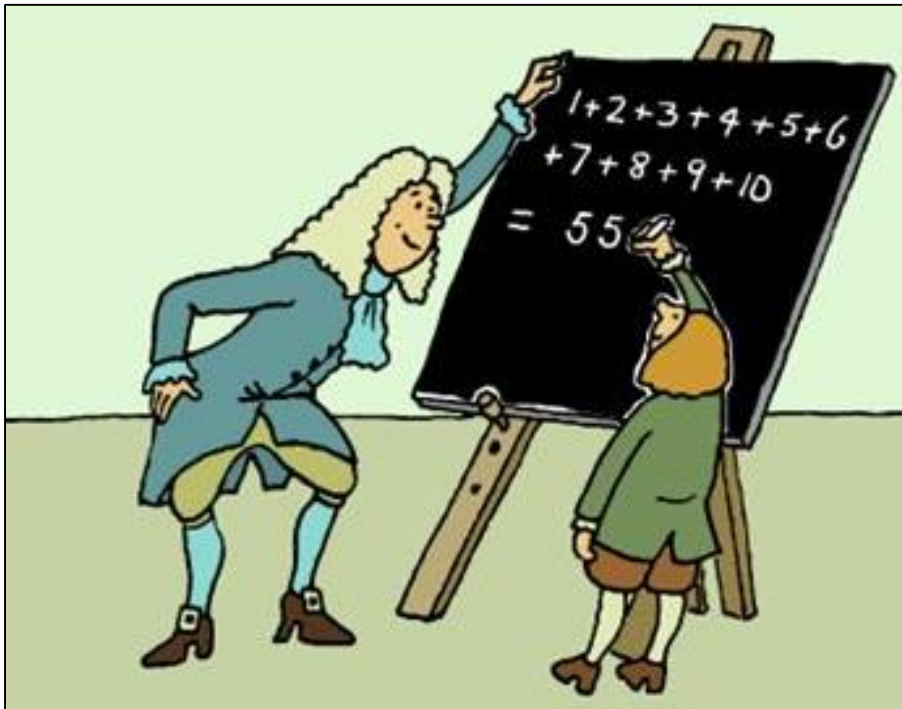
$$P(I) = F(b) - F(a)$$



Uniform Distribution:
p.d.f. & c.d.f.

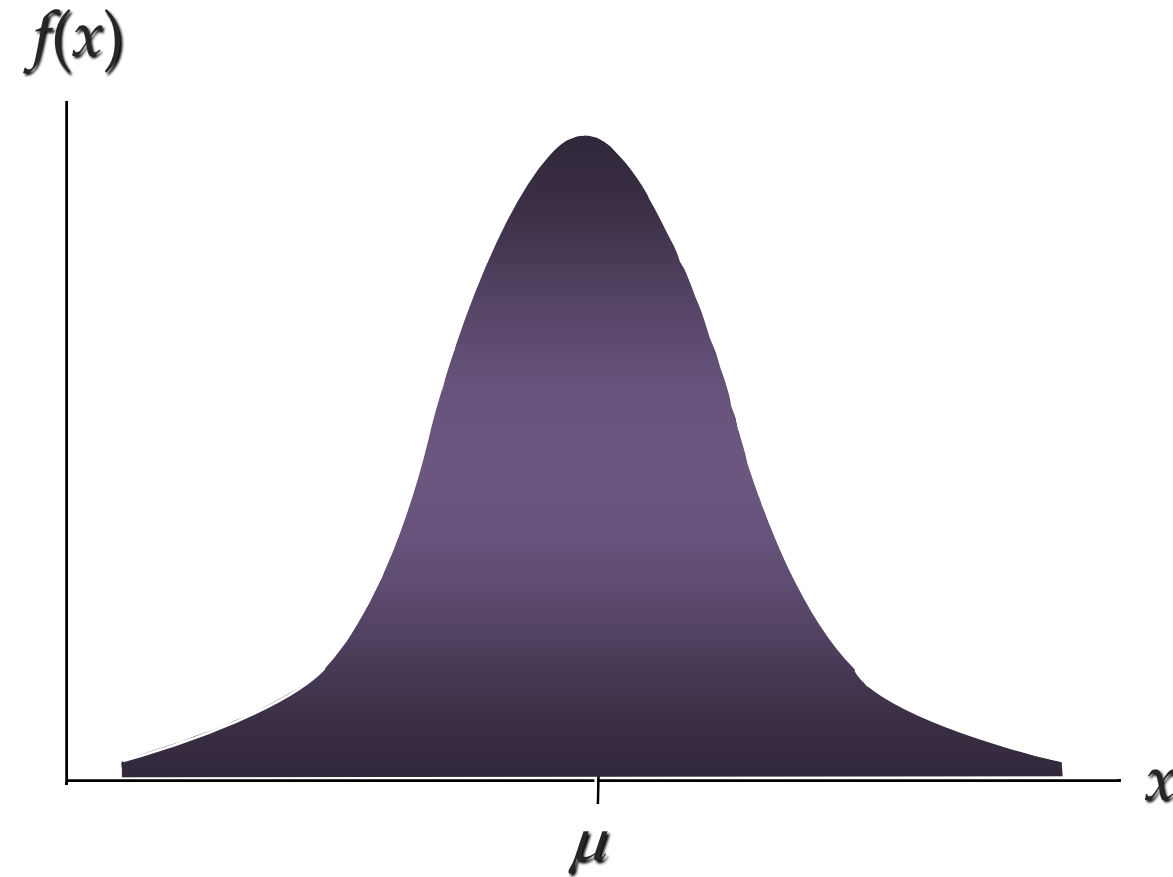
Gaussian distributions and the Central Limit Theorem

- ▶ Carl Freidreich Gauss, 1777-1855, Germany



Gaussian or Normal Probability Distributions

- ▶ The shape of the Gaussian, or Laplace-Gauss, or normal curve is often referred to as a bell-shaped curve
- ▶ The highest point on the normal curve is at the mean, which is also the median (and mode) of the distribution
- ▶ The normal curve is symmetric
- ▶ The standard deviation determines the width of the curve
- ▶ The total area under the curve is 1
- ▶ Probabilities for the normal random variable are given by areas under the curve



The normal density function

- ▶ Density functions for Gaussian r.v.s:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- ▶ We then say that the r.v X is normally distributed with mean μ and standard deviation σ
- ▶ We write $X \sim N(\mu, \sigma)$
- ▶ A random variable that has a normal distribution with
 - ▶ $\mu = 0$ and $\sigma = 1$
 - ▶ is called Standard Normal
 - ▶ The density function then becomes:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Randomistan electric bills

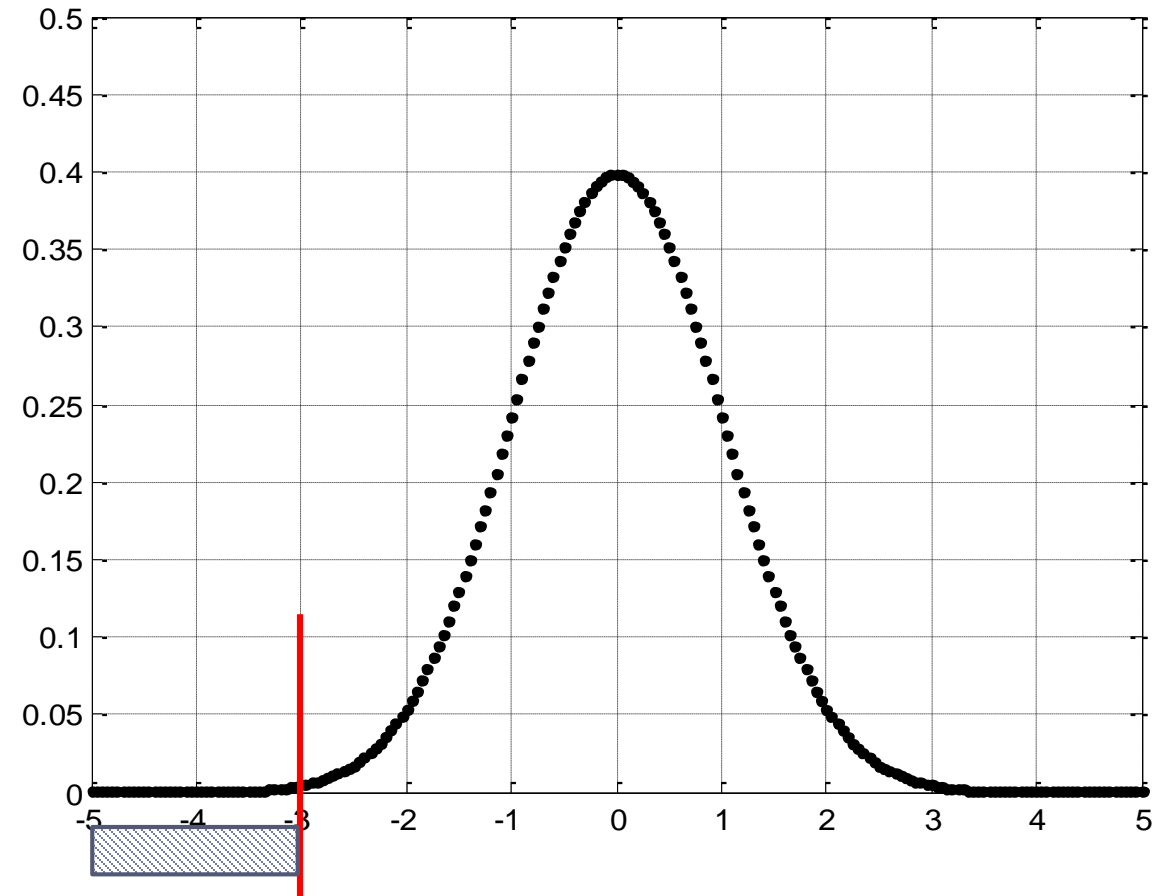
- ▶ Historical data shows that monthly electric bills for Stochastic Heights households are normally distributed with a mean of 225RSU and a standard deviation of 55RSU
- ▶ To encourage savings in electricity the mayor decided that starting in 2019 any household consuming less than 60RSU will be exempt from payment
- ▶ SH is expected to have 18K households in 2019
 - ▶ How many monthly bills do we expect to be exempt from payment in 2019?
 - ▶ Can you estimate the cost of this policy to the city budget?

A useful fact

- ▶ If $X \sim N(\mu, \sigma)$
then
- ▶ $Z = (X - \mu)/\sigma$ is a standard normal random variable
- ▶ You can think of Z as measuring, for every instance of X drawn, the distance of the obtained value, from the expected value, in units of standard deviations

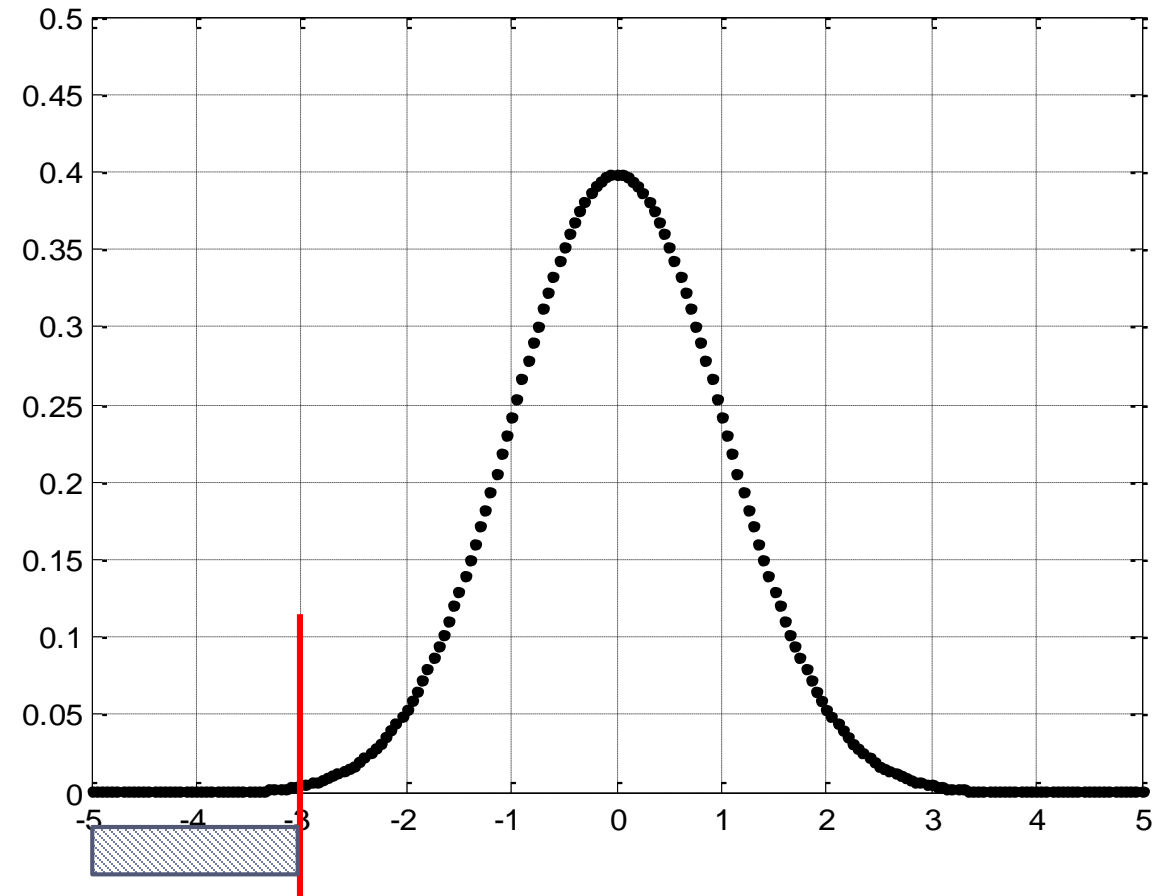
Back to SH...

- ▶ $z = (x - \mu) / \sigma$
- ▶ $= (60 - 225) / 55$
- ▶ $= -3$
- ▶ So – we want the area under the graph, to the left of the red line
- ▶ We will then multiply it by the total expected number of bills in 2019, namely $18K * 12$



Back to SH...

- ▶ $P(\text{Exempt})$
 $= 1 - \text{normcdf}(z = (x - \mu)/\sigma)$
 $= 1 - \text{normcdf}(3) = 0.0013$
- ▶ so the mayor should expect
~280 exempt bills
- ▶ We can therefore bound the
expected total cost of the
mayor's policy by $280 \cdot 60\text{RSU}$.



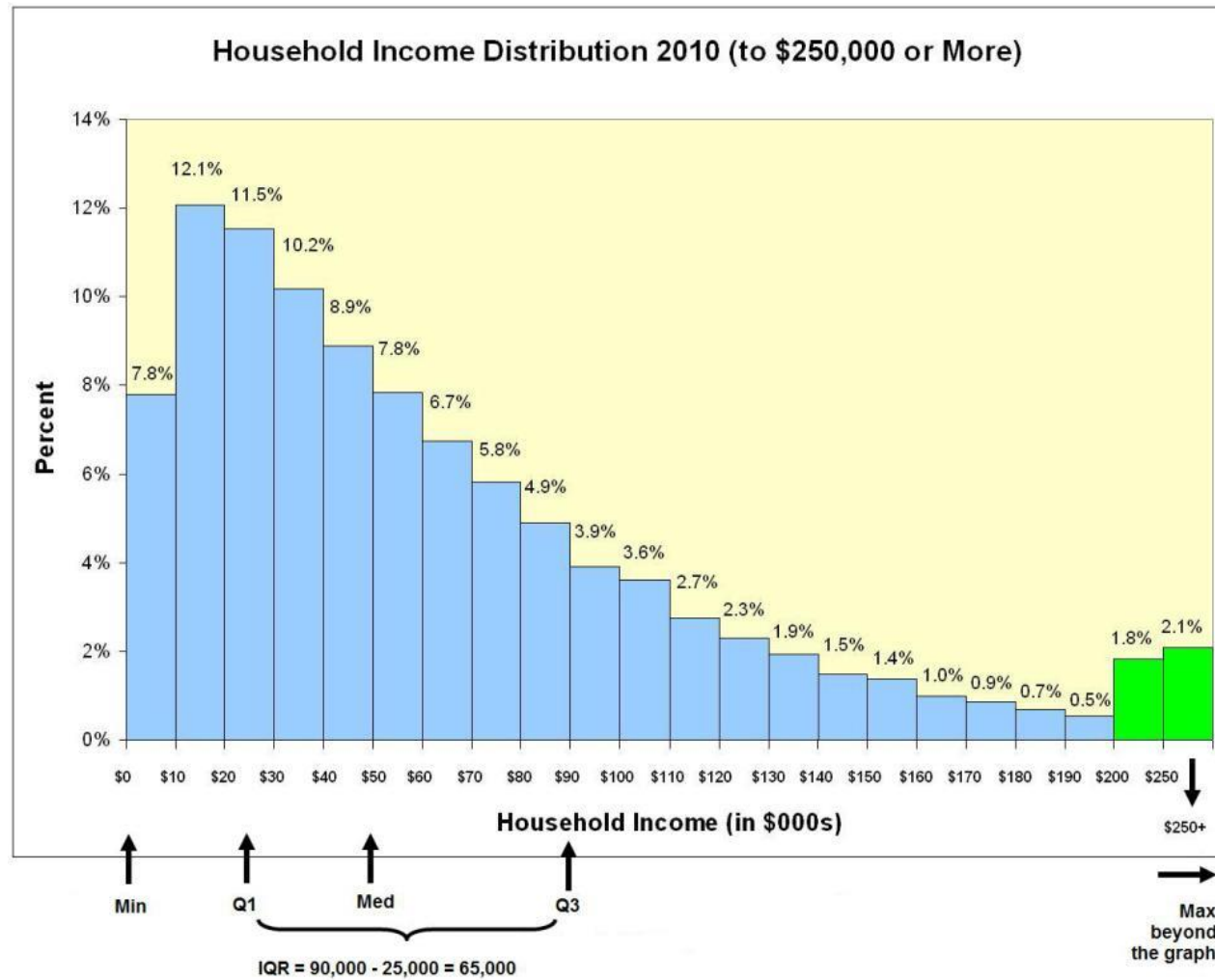
Intro to Central Limit Theorem

- ▶ A foundational assumption in the electricity bills story was that they determined a normal distribution for the monthly bills
- ▶ It enabled to use the normal distribution characteristics
- ▶ When can such a conclusion be drawn? How ubiquitous is the normal distribution?
- ▶ What can be done in more general cases?

Bad news - not all is really normal...

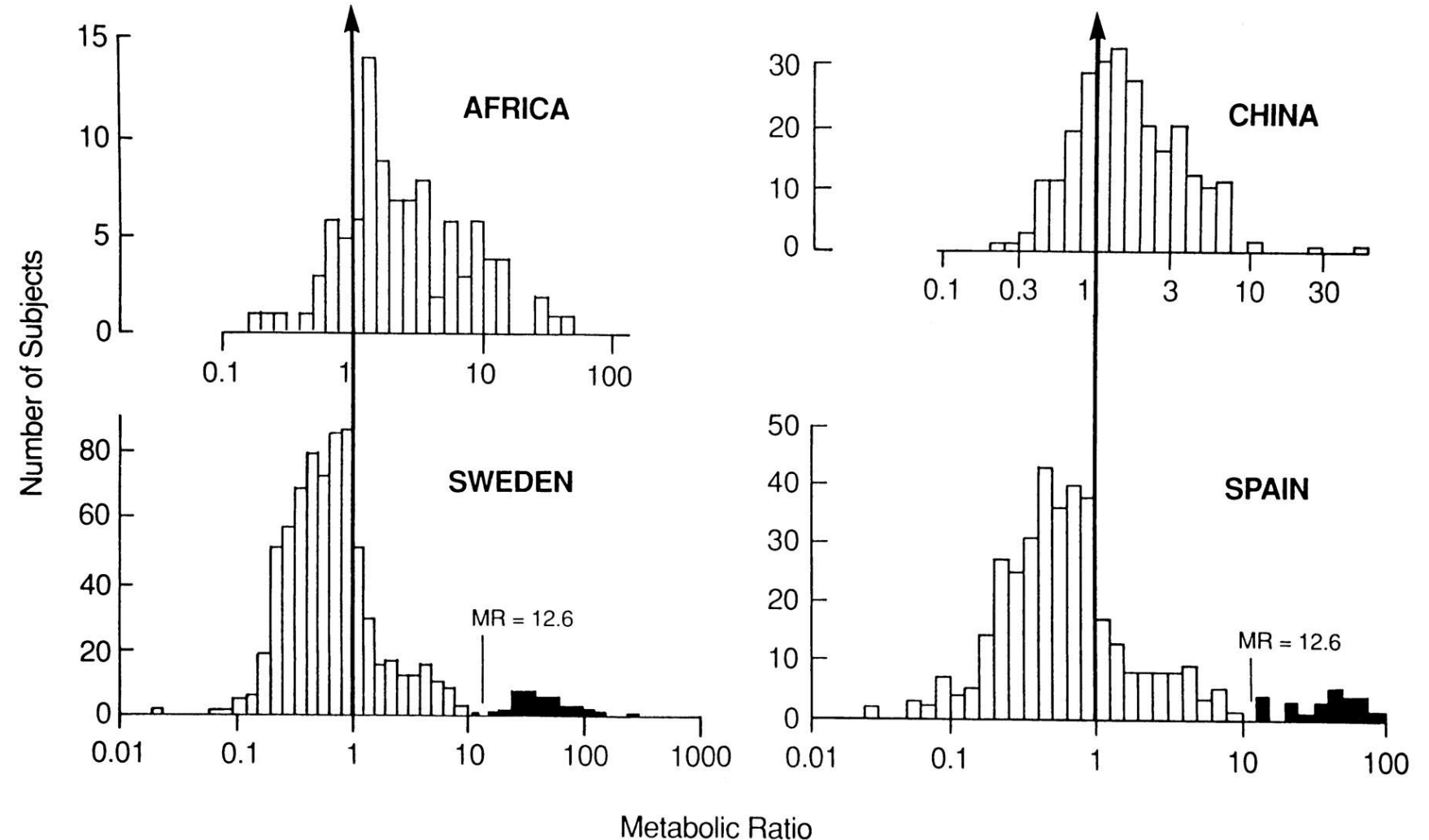


Example



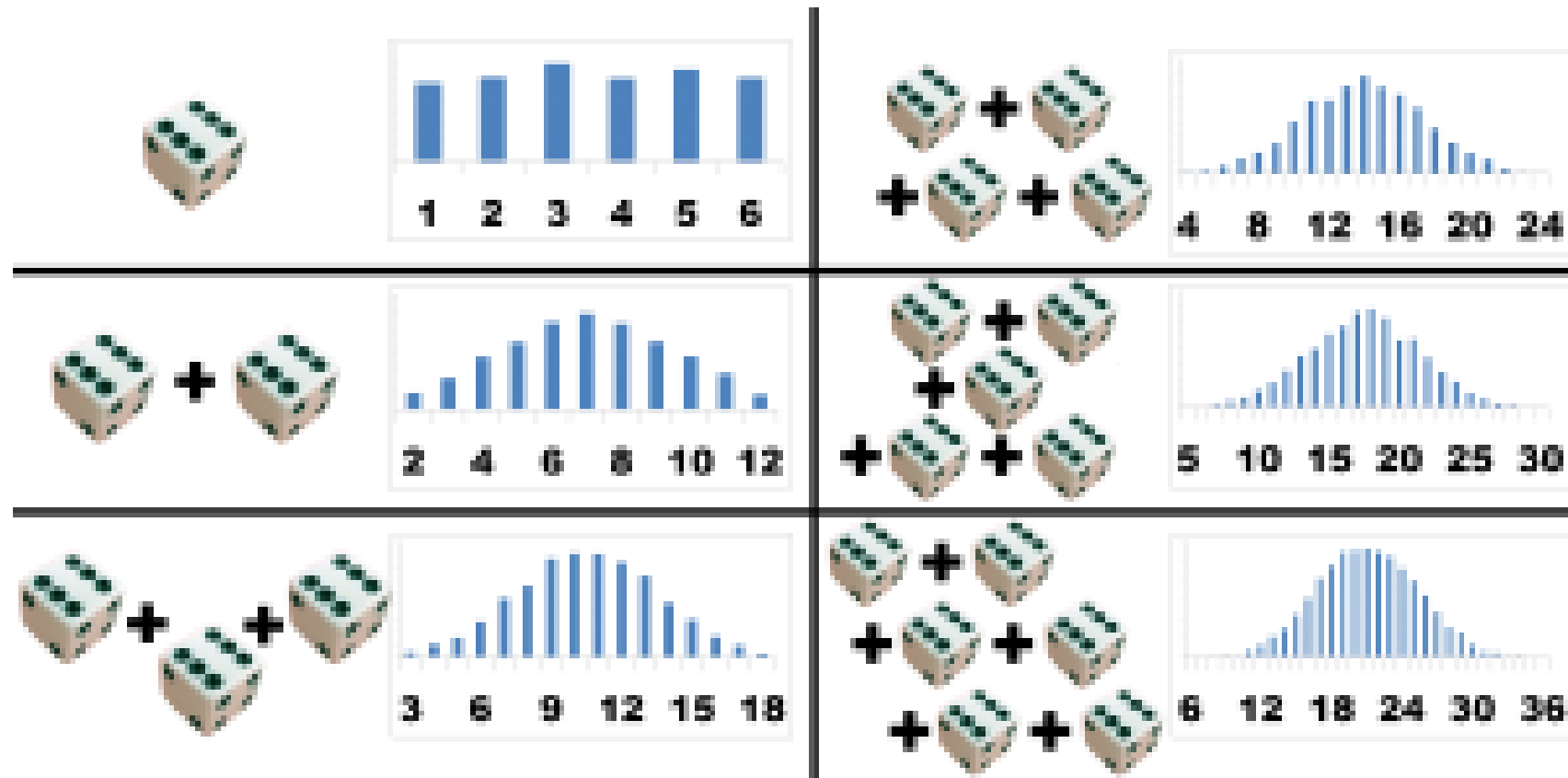
Example

- ▶ X axis: ratio of nonmetabolized (no OH) drug to metabolized (hydroxylated) drug, in urine samples
- ▶ W Kalow
Pharmacokinetics
1997



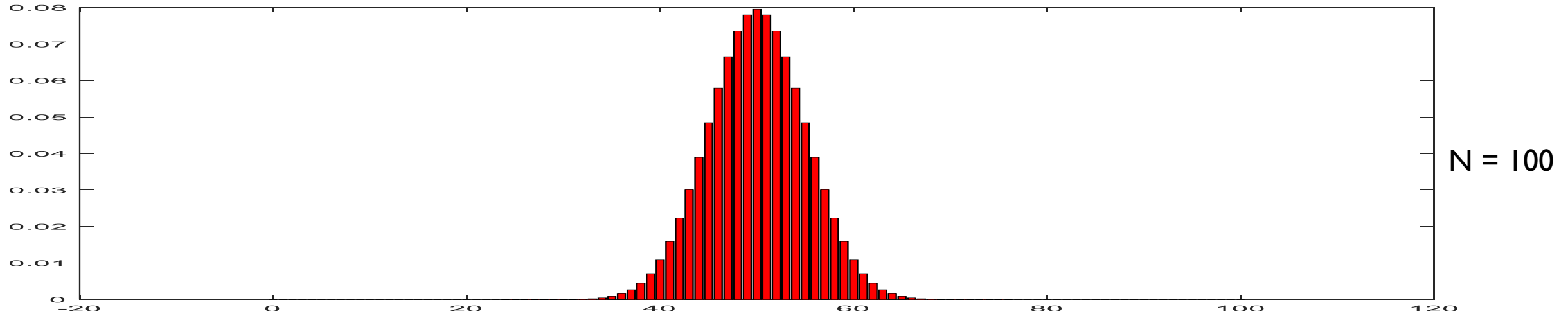
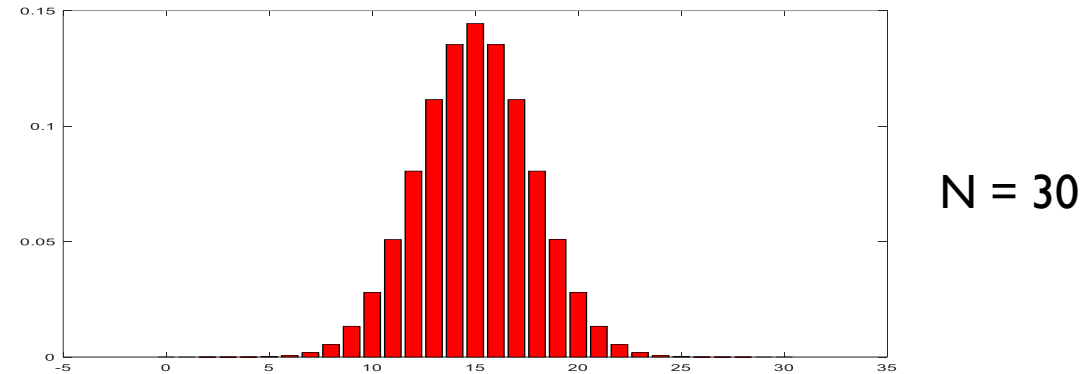
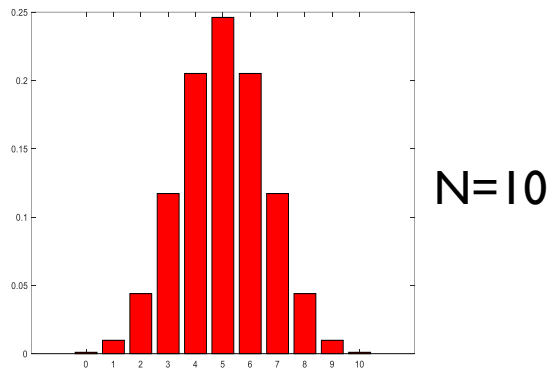
But...

- But most distributions average to be normal:
the Central Limit Theorem

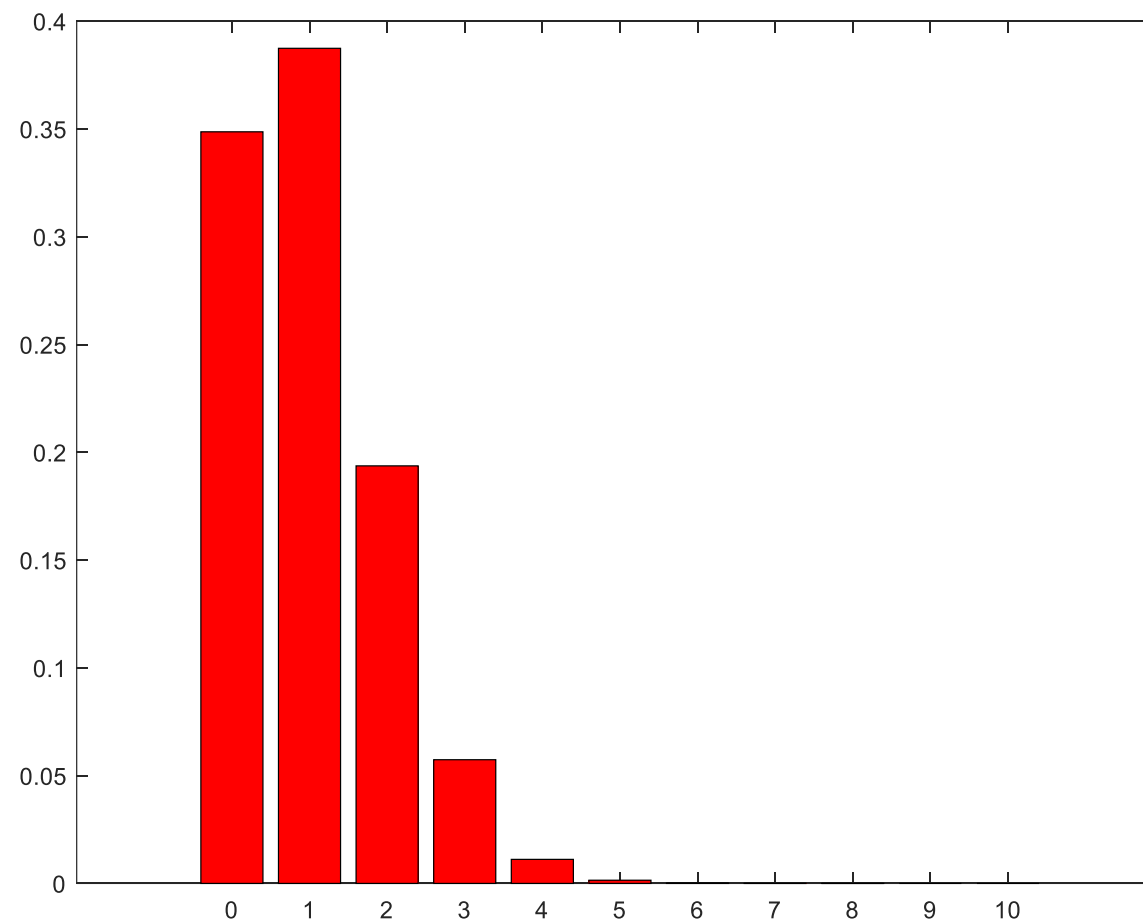


CLT

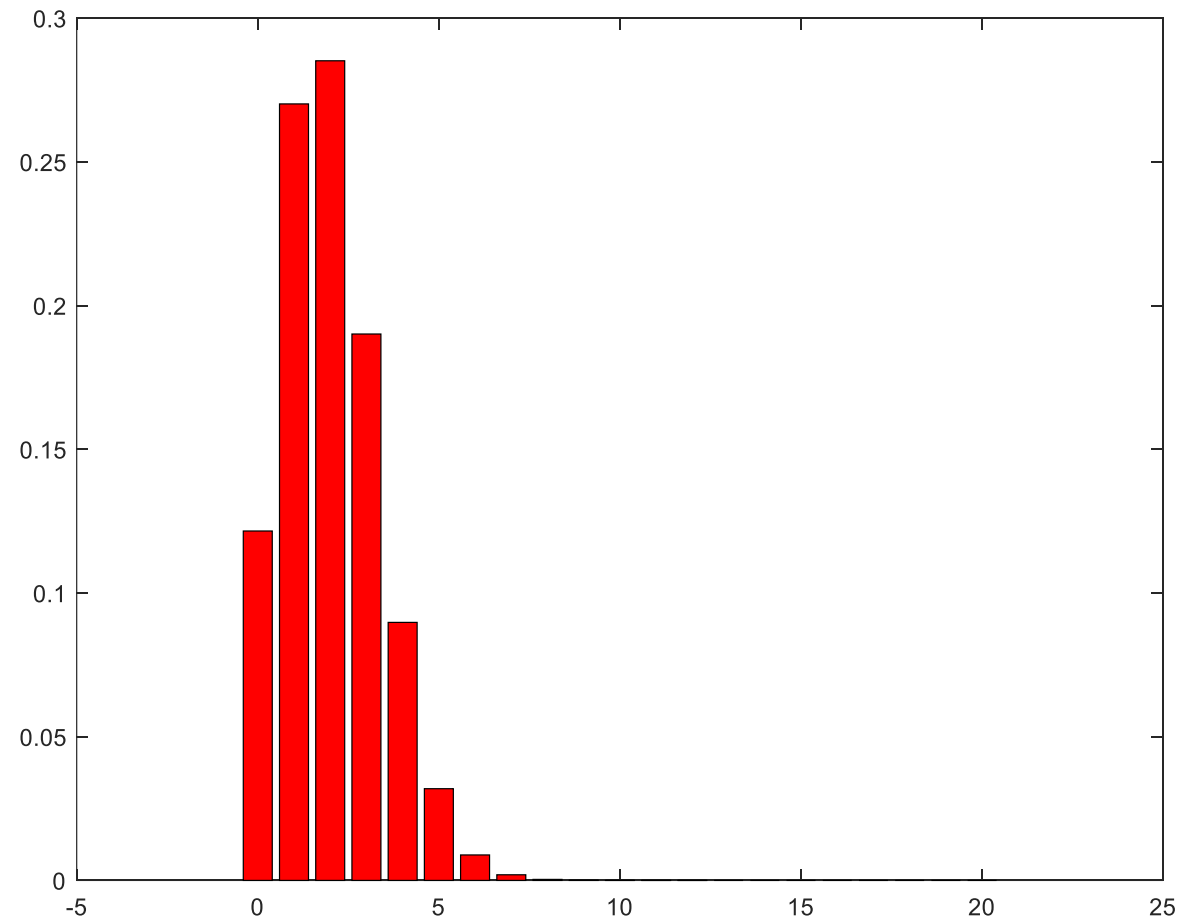
- ▶ The normal (Gaussian) distribution is important since it is the large sample limit behavior of (almost) all repeated sampling and averaging situations



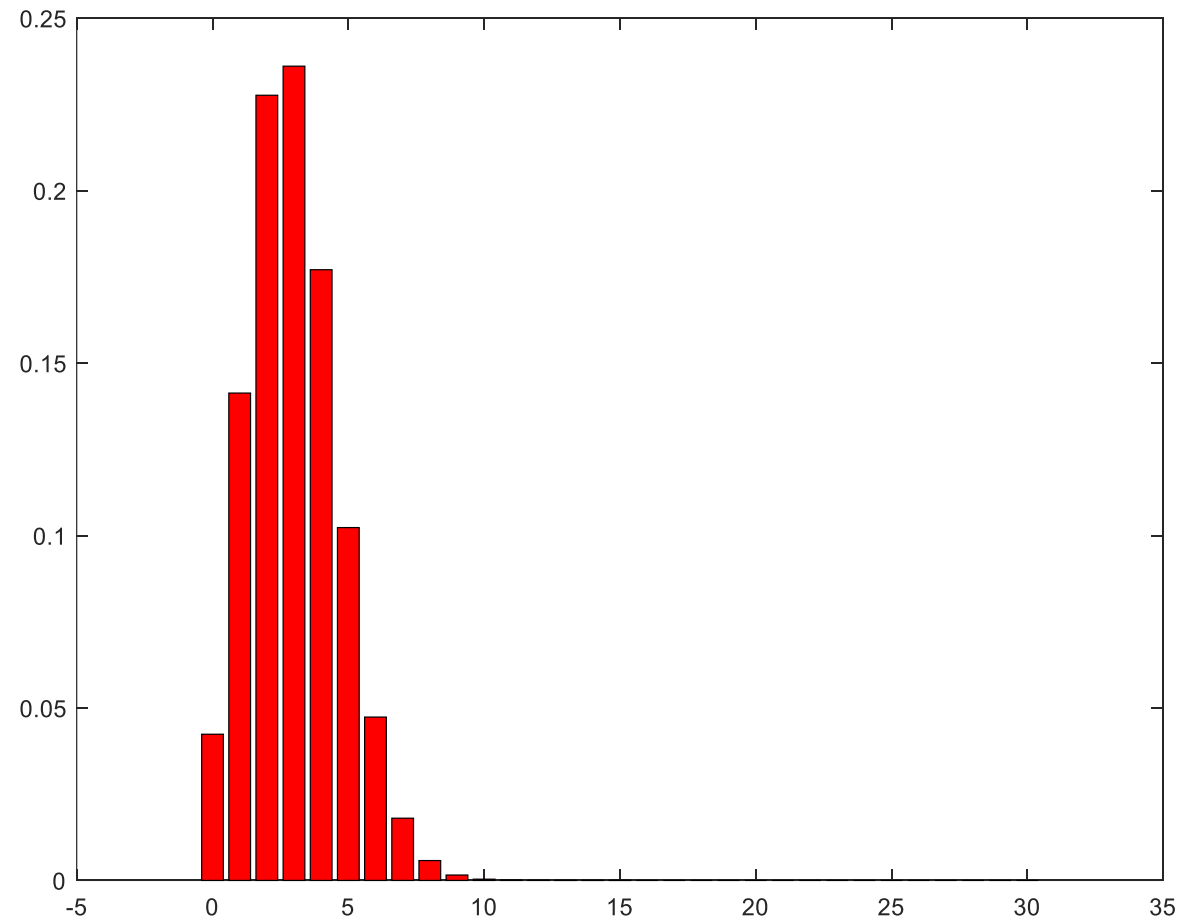
Binomials w $p=0.1$



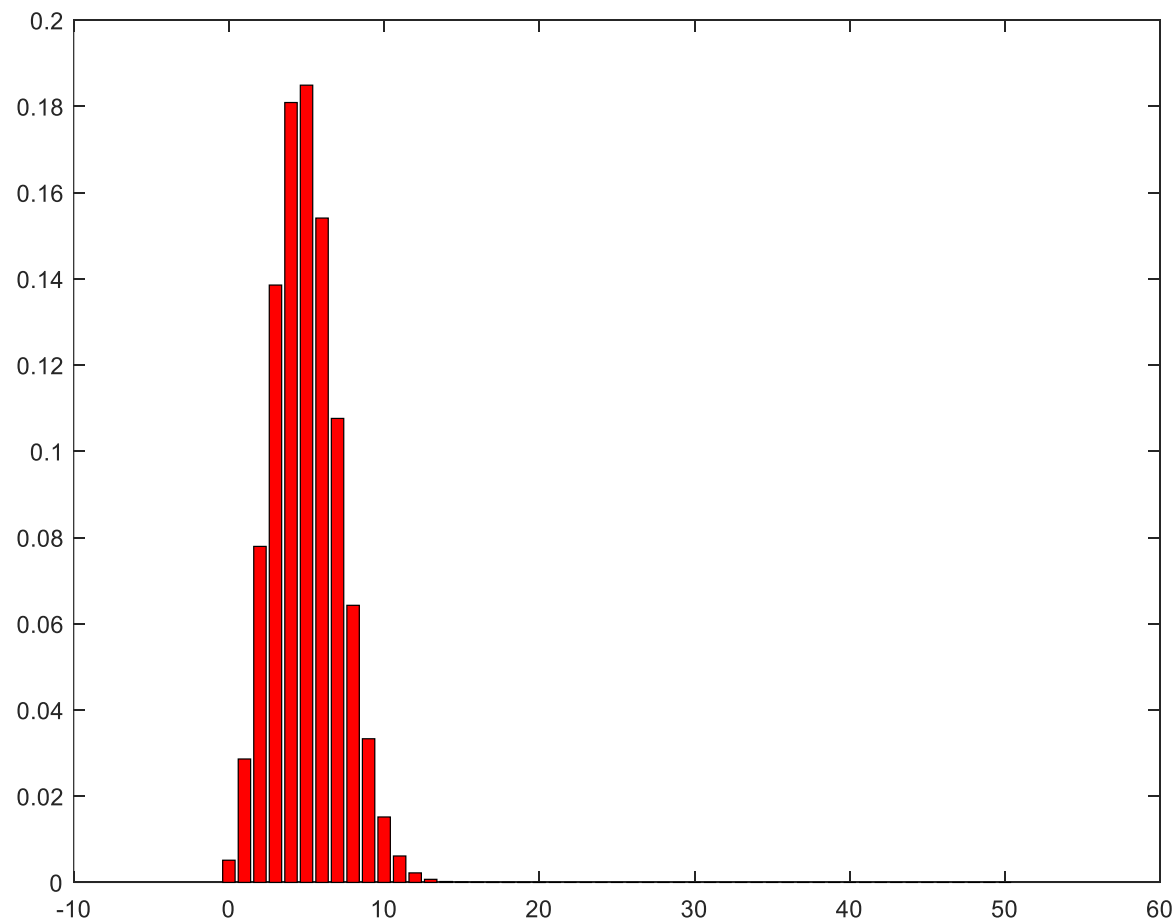
Binomials w $p=0.1$



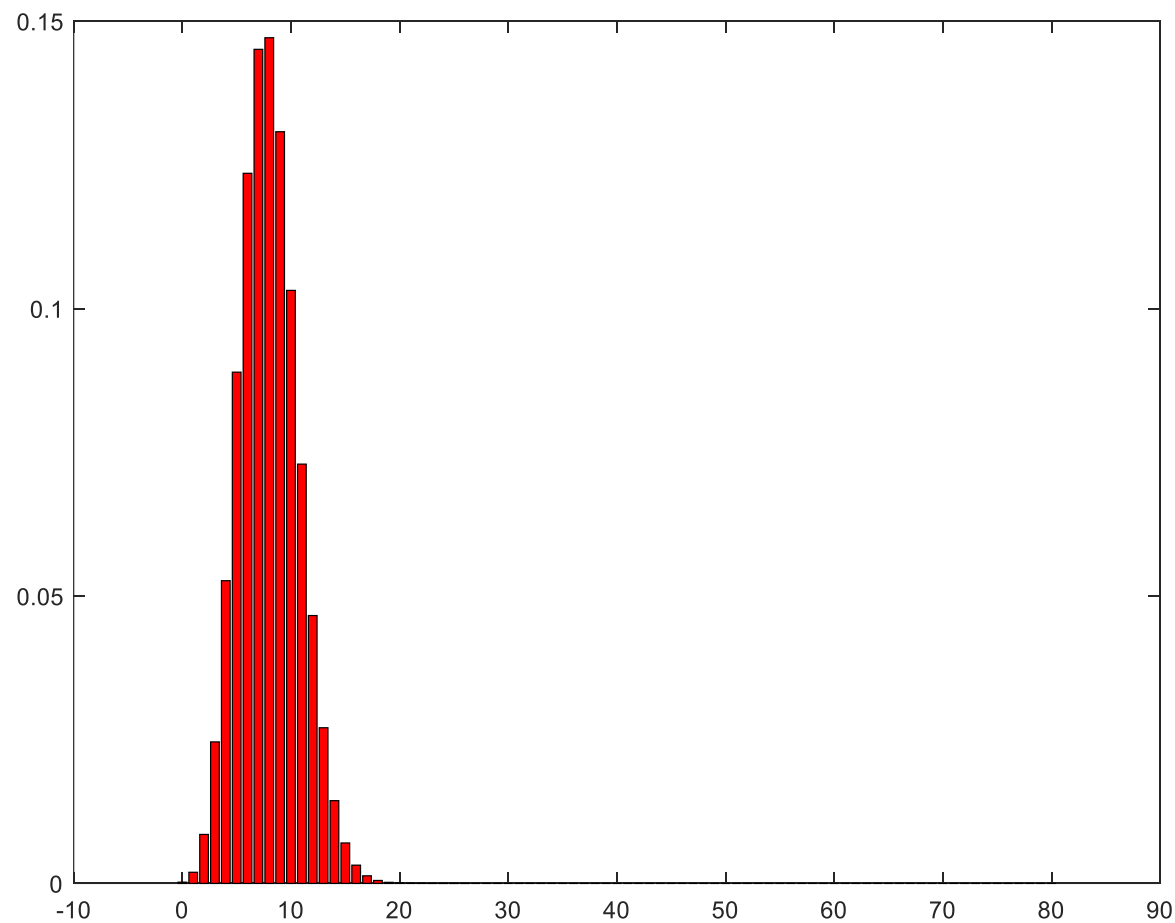
Binomials w $p=0.1$



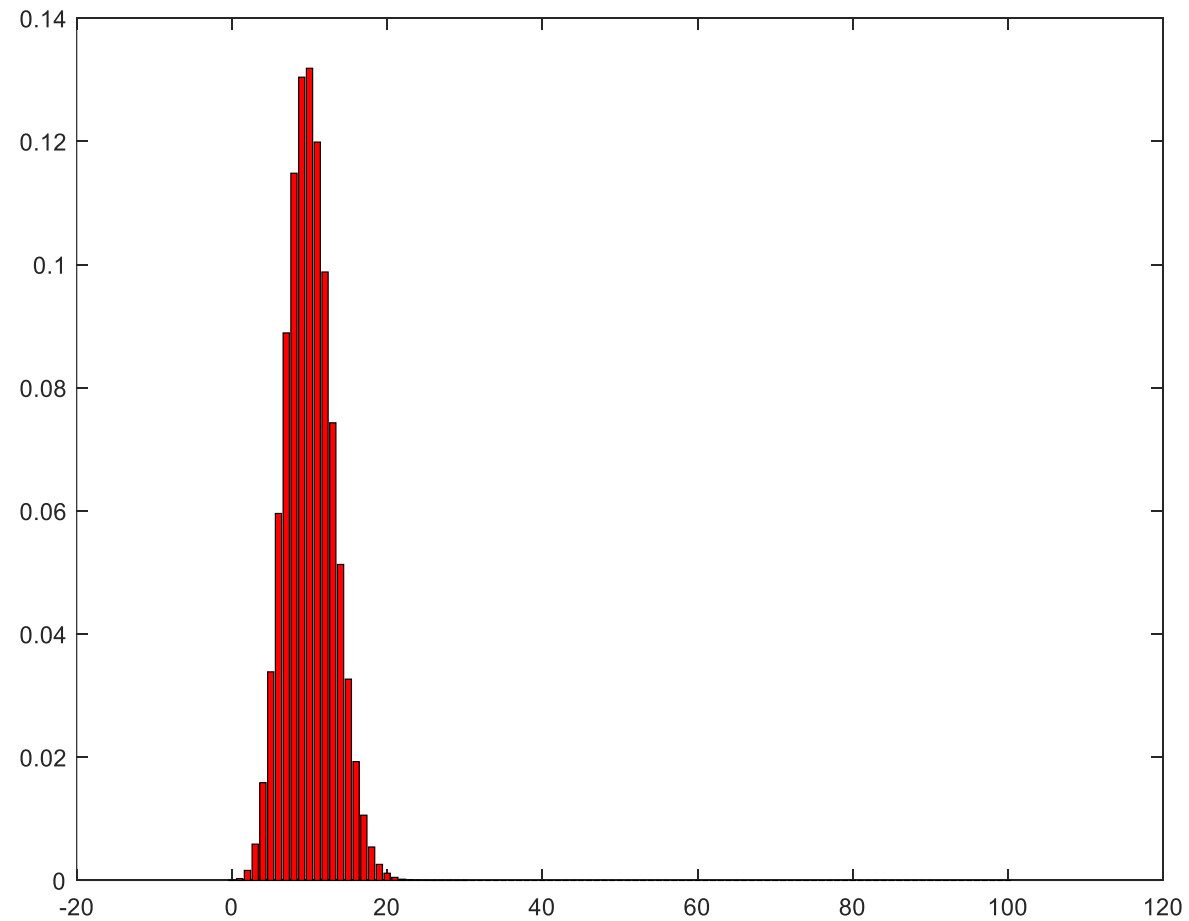
Binomials w $p=0.1$



Binomials w $p=0.1$



Binomials w $p=0.1$



The Central Limit Theorem

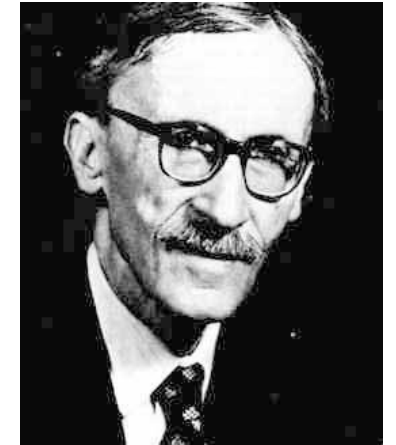
(Lindeberg and Levy 1920)

- ▶ Let $X_1, X_2, X_3, \dots, X_n$ be random variables all sampled **independently** from **the same distribution** with mean μ and (finite non 0) variance σ^2

- ▶ Let \bar{X}_n be the average of $X_1, X_2, X_3, \dots, X_n$

- ▶ Then for any fixed number x we have

$$\lim_{n \rightarrow \infty} P \left(\frac{\sqrt{n}}{\sigma} (\bar{X}_n - \mu) \leq x \right) = \Phi(x)$$



- ▶ where $\Phi(x)$ is the standard normal density function
- ▶ In simpler words – for (**almost**) any distribution, if we sample it (**sufficiently**) many times and then average then our distance from the mean, properly scaled, is (**basically**) standard normally distributed
- ▶ Alternatively - \bar{X}_n is approximately normally distributed with mean μ and variance σ^2/n



Exercise (11)

- ▶ In this exercise you will construct trajectories of dice rolling results in the following way:
 - ▶ The first roll, X_1 , is $\text{Unif}(1..6)$
 - ▶ After i rolls are determined the $i+1^{\text{st}}$, X_{i+1} , is drawn according to the row that corresponds to the value of X_i in the matrix T below

$$T = \begin{pmatrix} 0.5 & 0.25 & 0 & 0 & 0 & 0.25 \\ 0.25 & 0.5 & 0.25 & 0 & 0 & 0 \\ 0 & 0.25 & 0.5 & 0.25 & 0 & 0 \\ 0 & 0 & 0.25 & 0.5 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.5 & 0.25 \\ 0.25 & 0 & 0 & 0 & 0.25 & 0.5 \end{pmatrix}$$

- ▶ That is, if the 10th roll was 4 then the 11th one will be 0.25 on 3, 0.5 on 4 and 0.25 on 5
- ▶ This mechanism of generating trajectories is a Markov chain. The matrix T is the transition matrix for this chain. The chain here starts with the uniform distribution as the initial distribution

Exercise (11)

- ▶ Construct 1000 trajectories, each of length 20
 - ▶ What do you expect the average value of all 20 numbers in a trajectory to be?
 - ▶ Compute the average value of each such trajectory. Draw a histogram of the 1000 numbers you received, using 100 bins
 - ▶ What does the distribution look like? What are the empirical mean and the std?
- ▶ Construct 1000 trajectories, each of length 2000
 - ▶ What do you expect the average value of all 2000 numbers in a trajectory to be?
 - ▶ Compute the average value of each such trajectory. Draw a histogram of the 1000 numbers you received, using 100 bins
 - ▶ What does the distribution look like? What are the empirical mean and the std?
- ▶ Draw normal fit curves on your two histograms

Gaussian mixtures

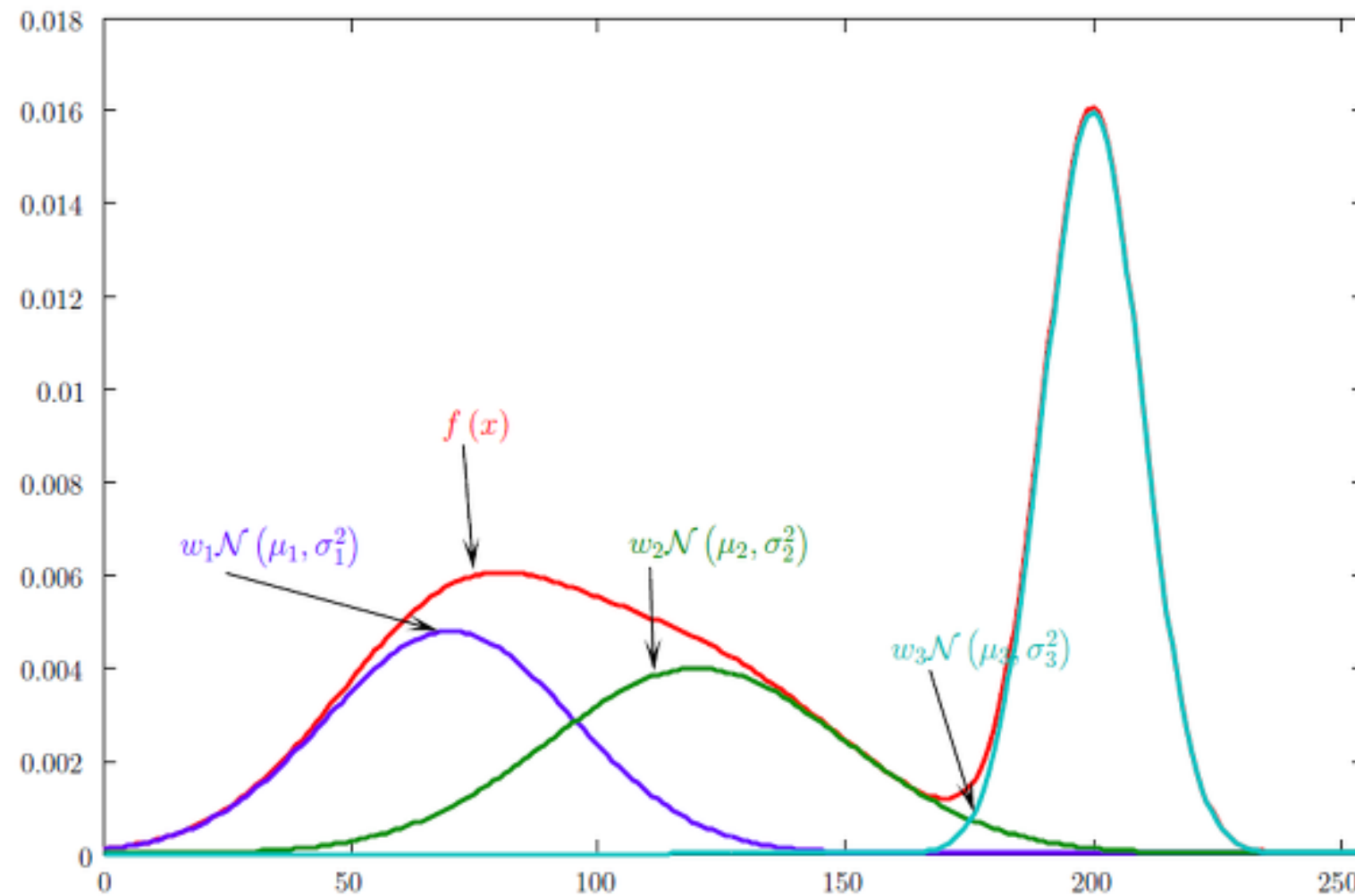
- ▶ X is a Gaussian Mixture RV if the density function for X's distribution is:

$$f(x) = \sum_{i=1}^k w_i f_i(x)$$

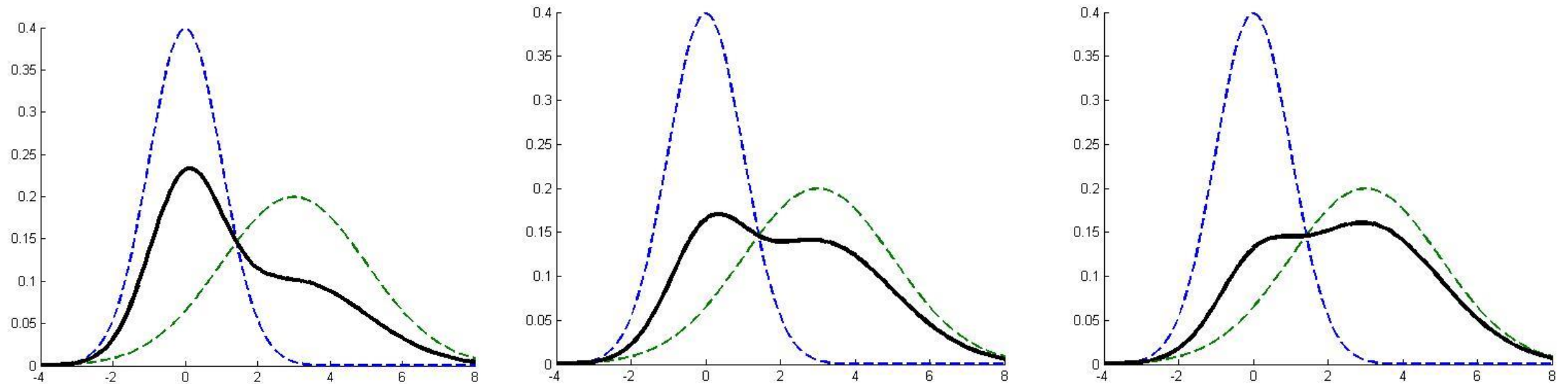
- ▶ Where each one of the f_i density functions is a Gaussian density:

$$f_i(x) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-(x-\mu_i)^2 / 2\sigma^2}$$

What does f look like?



GMD dependence on weights

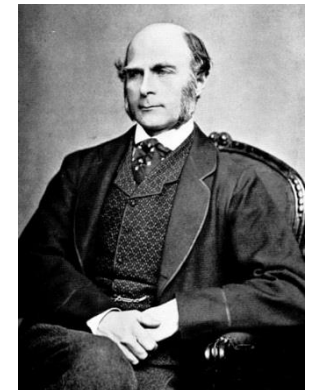
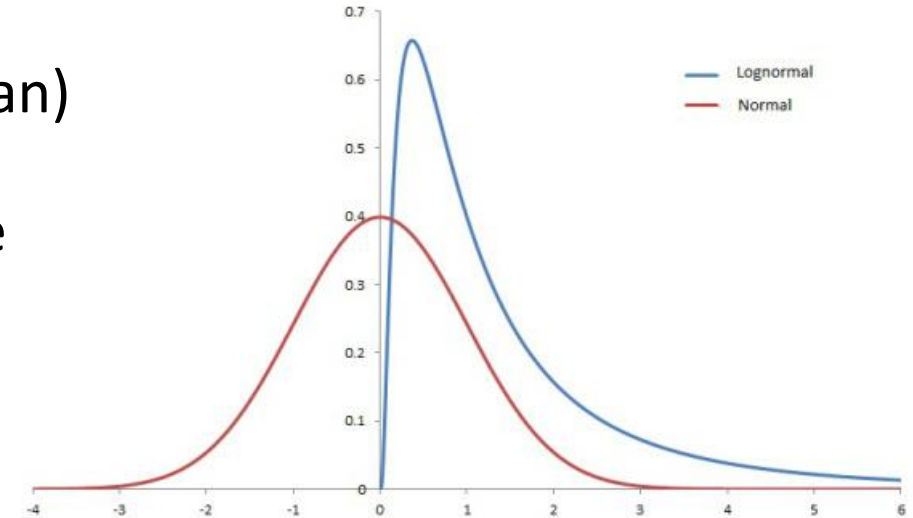


Exercise (12)

- ▶ In this question $N(\mu, \sigma^2)$ stands for a normal distribution with mean μ and variance σ^2
- ▶ Fred, Mel and Sid are repair technicians who work for Randobezeq – a phone company
- ▶ Fast Fred takes time which is distributed as $N(30, 25)$ to repair a telephone line failure at a customer's home
- ▶ Medium Mel takes time which is $N(35, 49)$ for the same task
- ▶ Slow Sid takes time which is $N(40, 100)$ for the same task
 - ▶ Fred is due to arrive to repair your phone at 11AM tomorrow. How confident can you be that you will be done by 11:45?
- ▶ When a customer in North Randomistan orders a repair, there is a 50% chance Fred will do the work and 25% each that Mel or Sid will do the work
 - ▶ What is the distribution of the duration of repair in North Randomistan?
 - ▶ Let Φ denote the CDF of a standard normal random variable. Use Φ to express the CDF of the duration of a repair in North Randomistan.
 - ▶ If the repair starts at 11AM, what is the earliest time for which the customer can assume, at a 95% certainty, that the repair will be already done?

Log-normal (Galton) distributions

- ▶ A random variable Y is said to have a log-normal distribution if its log, $\log(Y)$, has a normal (Gaussian) distribution
- ▶ In other words – Y is log-normal if $Y = e^X$ for some Gaussian X
That is: $Y = e^{\mu + \sigma Z}$, where Z is standard normal
- ▶ Lognormals are always positive
- ▶ Can be useful in modelling intrinsically positive quantities
- ▶ Mean, mode and median are different from each other
- ▶ The log-normal distribution has a heavy right side tail
- ▶ μ and σ are called the location and scale of Y . They are NOT the mean and std of Y
- ▶ They are the mean and std of $X = \ln(Y)$



Francis Galton,
1822-1911,
British statistician

Log-normal (Galton) distributions

- ▶ Median:

$$e^{\mu}$$

- ▶ Mean:

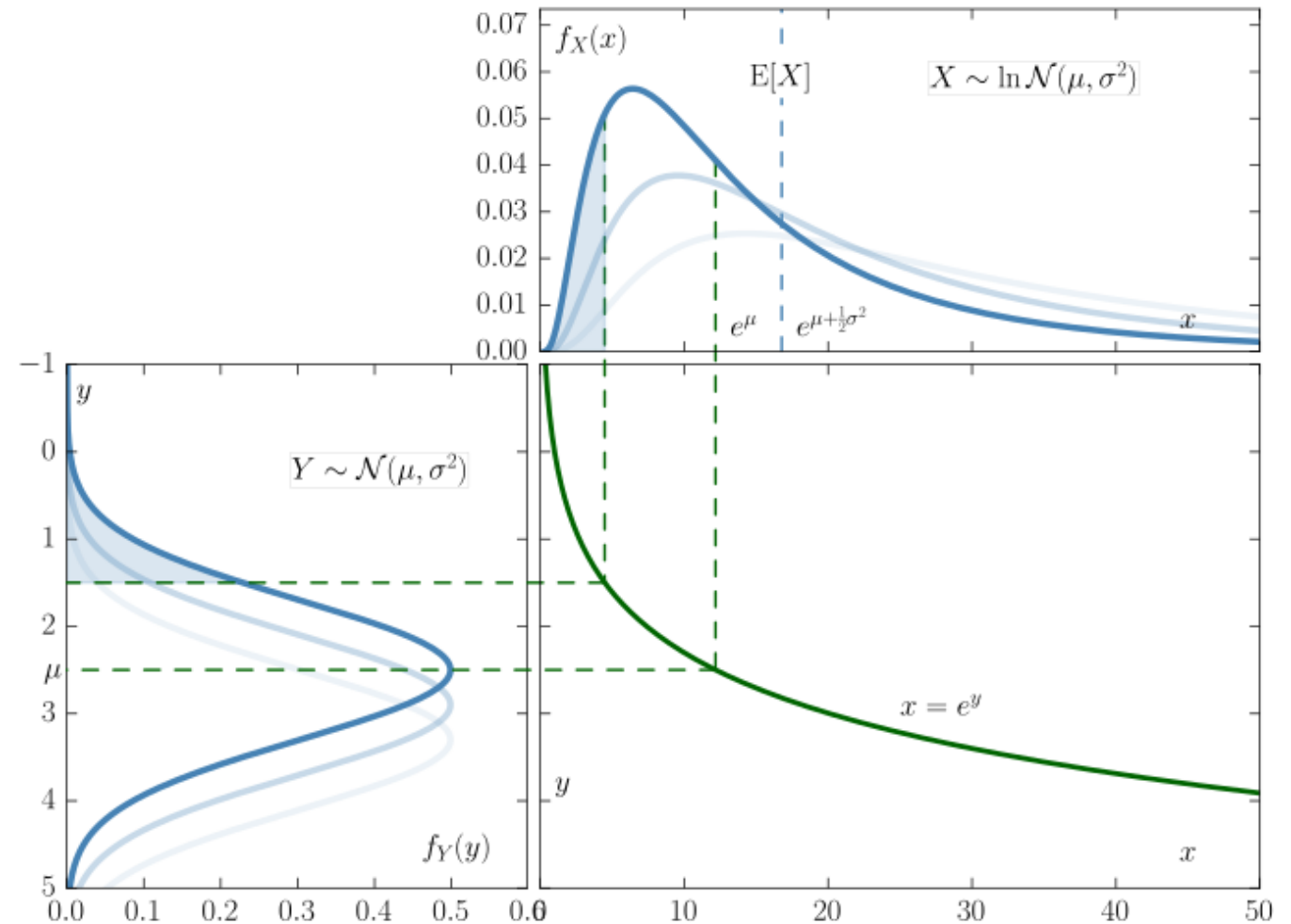
$$e^{\mu + \frac{\sigma^2}{2}}$$

- ▶ Mode:

$$e^{\mu - \sigma^2}$$

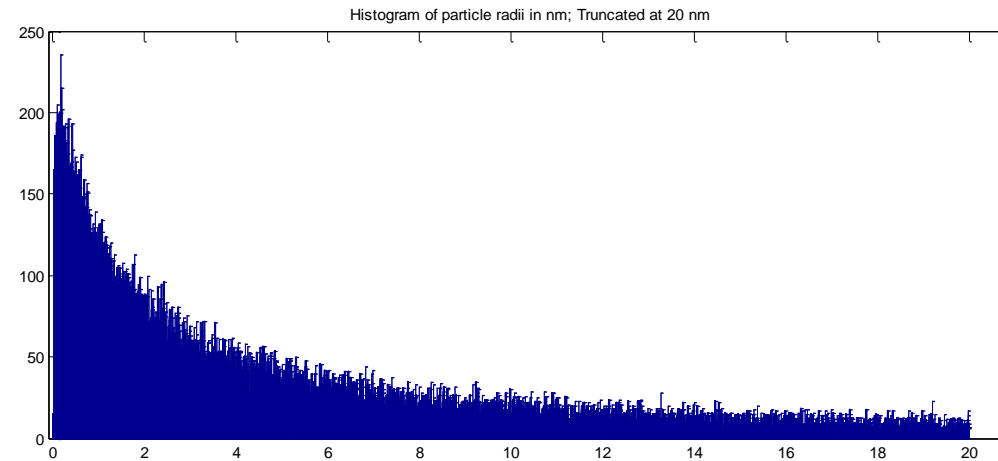
- ▶ Variance:

$$(e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$$



Exercise (13)

- ▶ A scientist is generating nanoparticles for an experiment. She observes the following distribution of particle radii, in nms (nano-meters):



- ▶ This histogram representation of the distribution is calculated from 100K particles
- ▶ The x-axis units are nms
- ▶ The histogram is truncated at 20 nm
- ▶ 30687 particles of the 100K measured had radius ≥ 20 nm

Exercise (13)

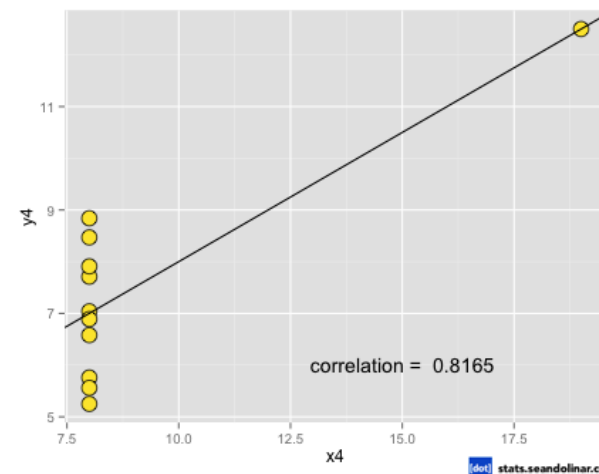
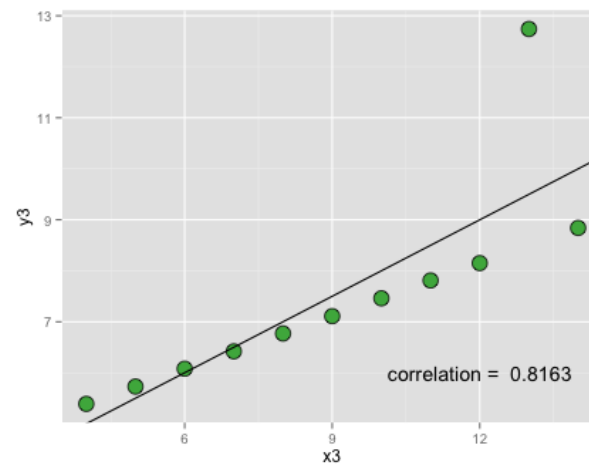
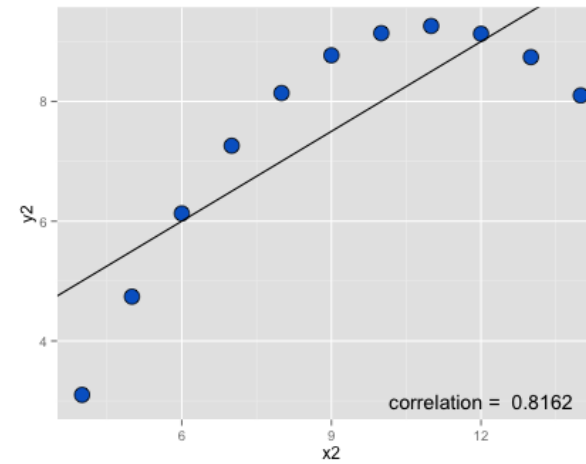
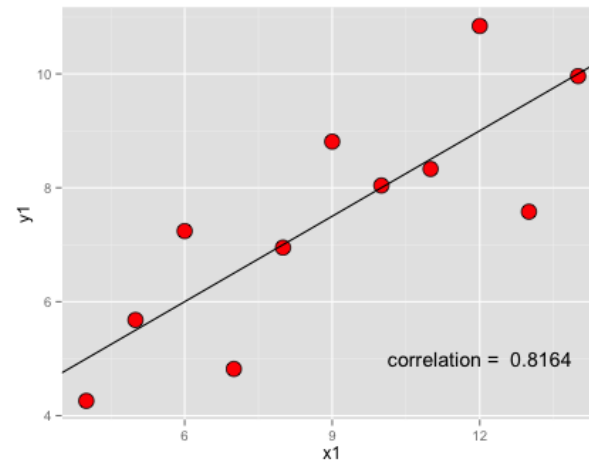
- ▶ For the above data representing 100K particles, the scientist calculated empirical statistics:

$$emp - mean = e^4 \text{ nm} \qquad emp - std = \sqrt{e^{12} - e^8} \text{ nm}$$

- ▶ The empirical median of the distribution is at e^2 nm
 - ▶ Let R denote the random variable that represents radii of the particles generated by the scientist.
- What do you think the distribution of R is? Explain your answer.
 - According to the model you have developed what is the radius r so that # of particles with radius $< r = 20000$? (leave answer in exp notation if necessary)
 - The experiment requires at most 10% of particles to have a radius larger than e^4 nm. Show, based on your model, that the population generated here is therefore not adequate for the experiment.
 - The scientist can treat the particles and decrease all particle radii.
 A reasonably priced process will lead to all radii decreasing exactly \sqrt{e} fold (a particle with radius r will have radius $r \cdot 1/\sqrt{e}$ after the treatment).
 A more expensive process will lead to all radii decreasing exactly e fold (a particle with radius r will have radius $r \cdot 1/e$ after the treatment).
 She consulted with her statistician colleague as to whether either of the treatments will solve the problem and specifically as to whether the less expensive one will do it.
 What advice would you give in this case? Show all your calculations.

Measures of correlation in data

Anscombe Quadrant -- Correlation Demonstration



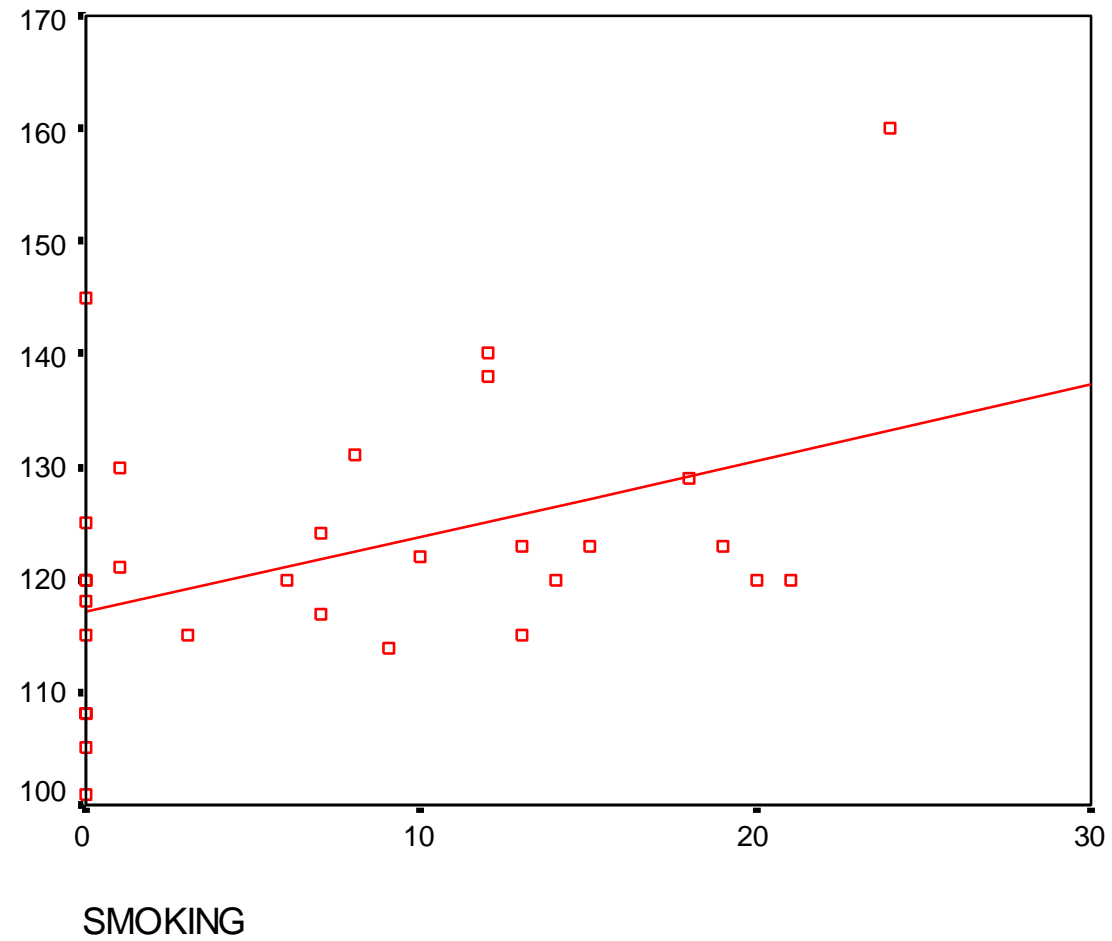
The Question

- ▶ Are two variables related?
 - ▶ Does one increase as the other increases?
 - ▶ e. g. skills and income,
expression of a TF and its target
 - ▶ Does one decrease as the other increases?
 - ▶ e. g. health problems and smoking,
expression of a miR and its target
- ▶ Correlation measures are mathematical tools that allow us to decide if a relationship exists and then get a numerical measure of how strong it is
- ▶ Caveat: no causality can be directly inferred

Assessing the significance of observed correlations

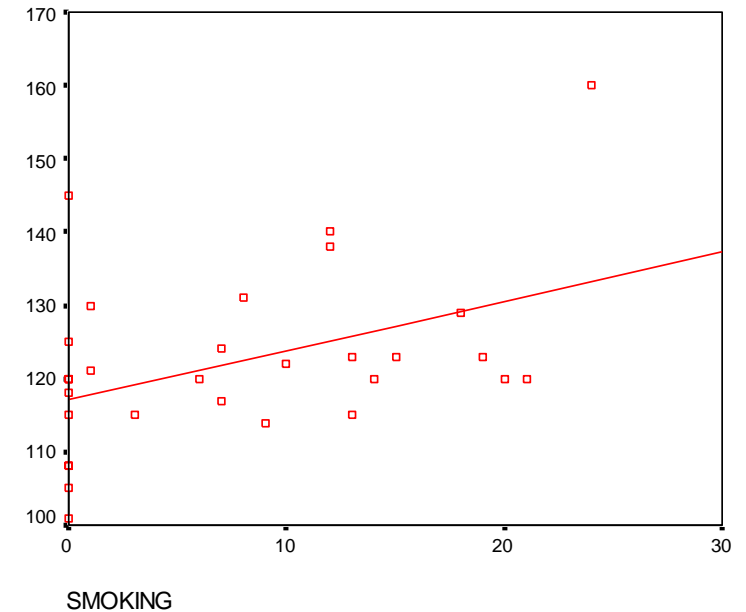
- ▶ Like any other finding resulting from analyzing data we will be interested in assessing the statistical significance of observations
- ▶ First, we will only define correlation measures
- ▶ Next, we will address statistical significance (p-values, CIs)

Trend?



Smoking and BP

- ▶ Note relationship is moderate, but (looks) real.
- ▶ Why do we care about relationship?
 - ▶ What would conclude if there were no relationship?
 - ▶ What if the relationship were near perfect?
 - ▶ What if the relationship were negative?
 - ▶ What if the relationship is strongly positive?
- ▶ High BP causes smoking?
- ▶ Smoking causes high BP?



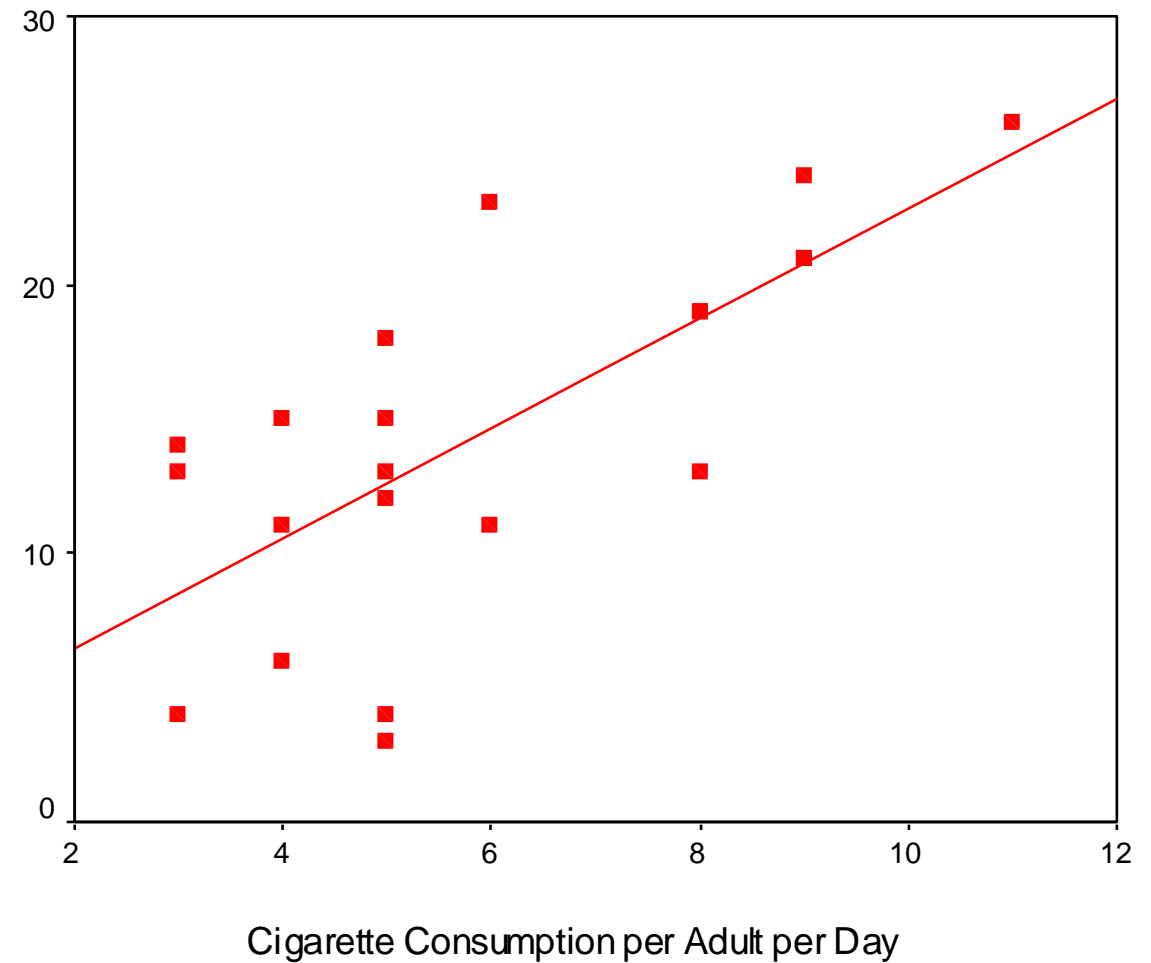
Heart Disease and Cigarettes

- ▶ Data on heart disease (mortality in 10K) and (average daily) cigarette smoking in 21 developed countries (Landwehr and Watkins, 1987)
- ▶ Data rounded for convenience

Country	Cigarettes	CHD
1	11	26
2	9	21
3	9	24
4	9	21
5	8	19
6	8	13
7	8	19
8	6	11
9	6	23
10	5	15
11	5	13
12	5	4
13	5	18
14	5	12
15	5	3
16	4	11
17	4	15
18	4	6
19	3	13
20	3	4
21	3	14

Is smoking related to CHD?

- ▶ As smoking increases, so does cor
- ▶ Relationship looks strong
- ▶ Not all data points on line.



Correlation measures – mathematical tools that address this question

- ▶ Mathematical correlation functions measure the relationship between two variables: co-relation
- ▶ There are many correlation functions, capturing different types of relationships
- ▶ We will study the mathematical properties for several important ones and discuss applications and examples

Covariance and Correlation Coefficient - reminder

- ▶ For two rvs X and Y defined on the same sample space:

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

- ▶ Alternatively:

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

- ▶ The correlation coefficient, for random variables, is defined as:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)}$$

Cauchy - Schwartz inequality

- ▶ For random variables U and V we have

$$[E(UV)]^2 \leq E(U^2)E(V^2)$$

- ▶ Setting $U = X - E(X)$ and $V = Y - E(Y)$ we get

$$[Cov(X, Y)]^2 \leq V(X)V(Y)$$

- ▶ And therefore

$$-1 \leq \rho(X, Y) \leq 1$$

Pearson correlation coefficient in observed data

$$\rho(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\text{sqrt}(\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2)}$$

Heart Disease and Cigarettes

Country	X (Cig.)	Y (CHD)	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X}) * (Y - \bar{Y})$
1	11	26	5.05	11.48	57.97
2	9	21	3.05	6.48	19.76
3	9	24	3.05	9.48	28.91
4	9	21	3.05	6.48	19.76
5	8	19	2.05	4.48	9.18
6	8	13	2.05	-1.52	-3.12
7	8	19	2.05	4.48	9.18
8	6	11	0.05	-3.52	-0.18
9	6	23	0.05	8.48	0.42
10	5	15	-0.95	0.48	-0.46
11	5	13	-0.95	-1.52	1.44
12	5	4	-0.95	-10.52	9.99
13	5	18	-0.95	3.48	-3.31
14	5	12	-0.95	-2.52	2.39
15	5	3	-0.95	-11.52	10.94
16	4	11	-1.95	-3.52	6.86
17	4	15	-1.95	0.48	-0.94
18	4	6	-1.95	-8.52	16.61
19	3	13	-2.95	-1.52	4.48
20	3	4	-2.95	-10.52	31.03
21	3	14	-2.95	-0.52	1.53

Empirical

Mean 5.95 14.52
SD 2.33 6.69
Sum

222.44

$$\sum (x_i - \mu_x)(y_i - \mu_y)$$

Empirical covariance

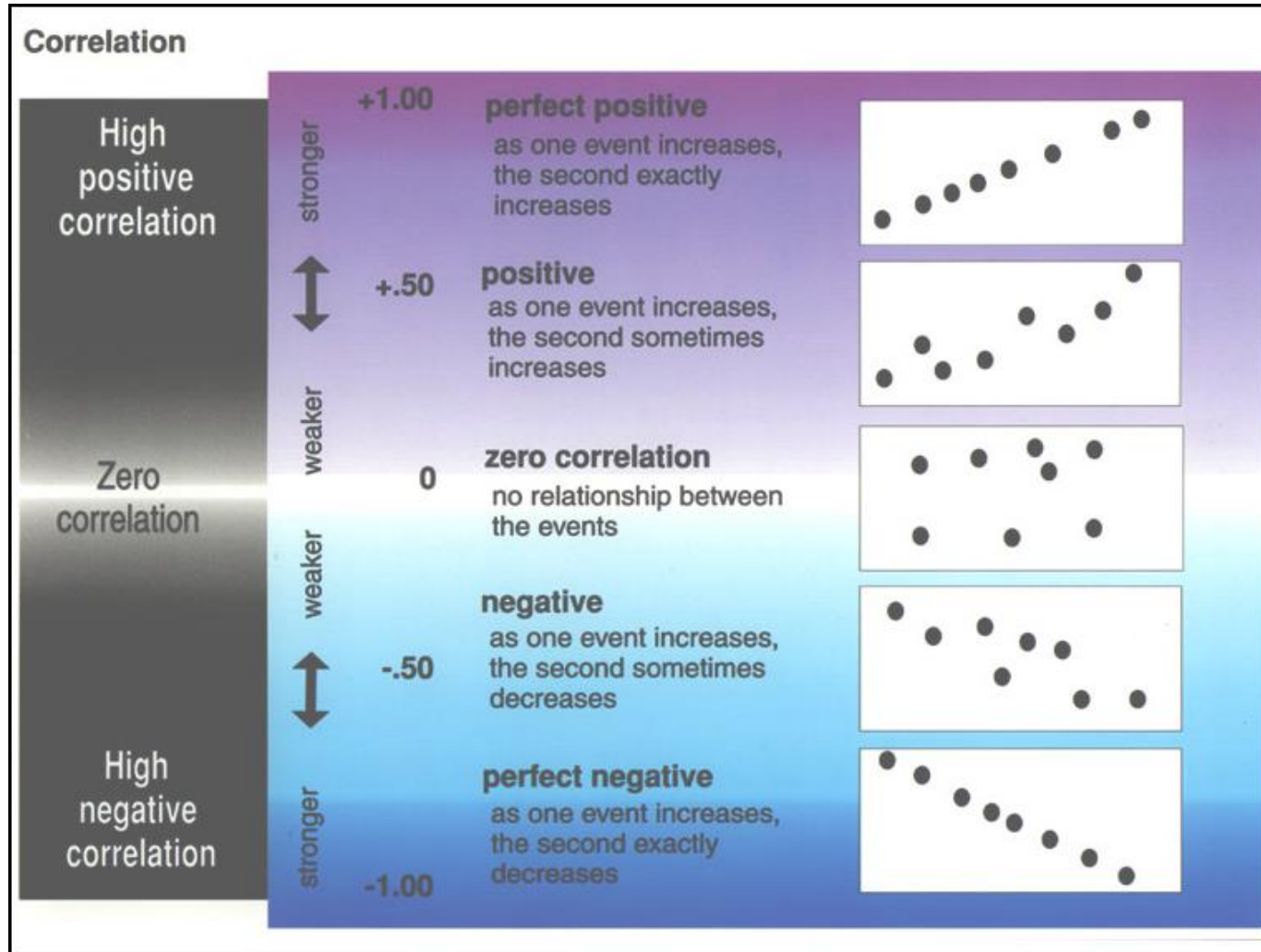
$$Cov_{cig.&CHD} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{N - 1} = \frac{222.44}{21 - 1} = 11.12$$

The Pearson correlation observed in the data

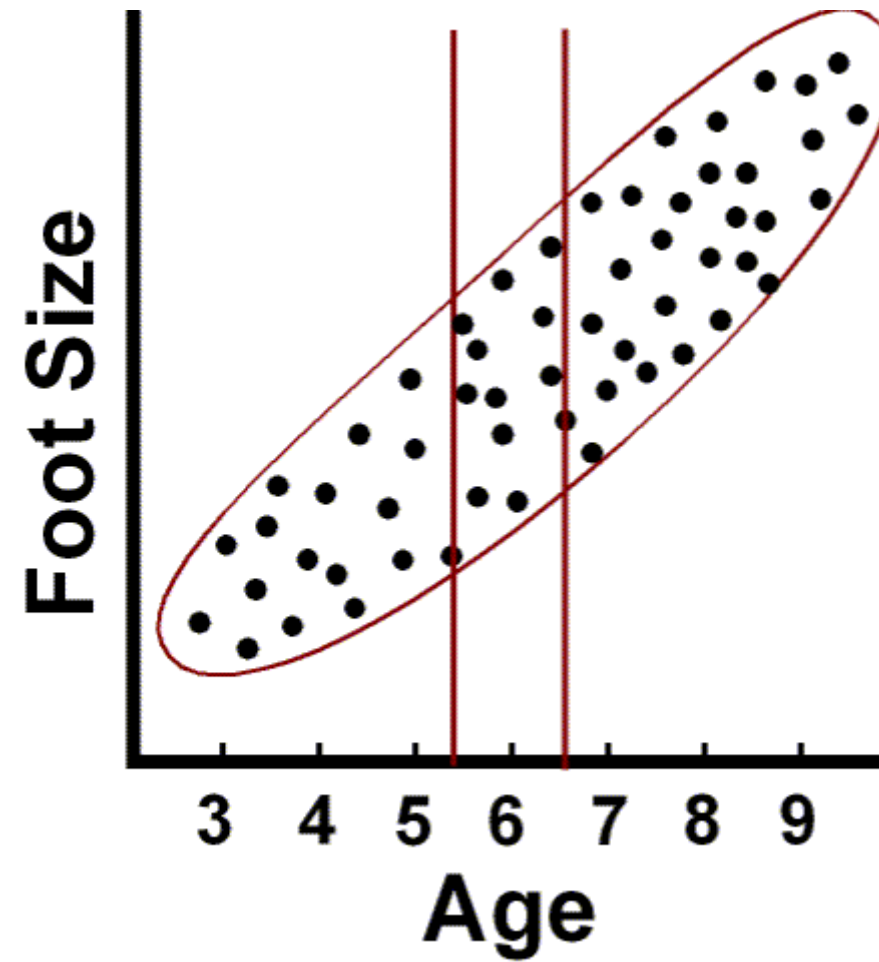
- ▶ $\text{Cov}_{XY} = 11.12$
- ▶ $s_X = 2.33$
- ▶ $s_Y = 6.69$

$$r = \frac{\text{COV}_{XY}}{s_X s_Y} = \frac{11.12}{(2.33)(6.69)} = \frac{11.12}{15.59} = .713$$

Correlation



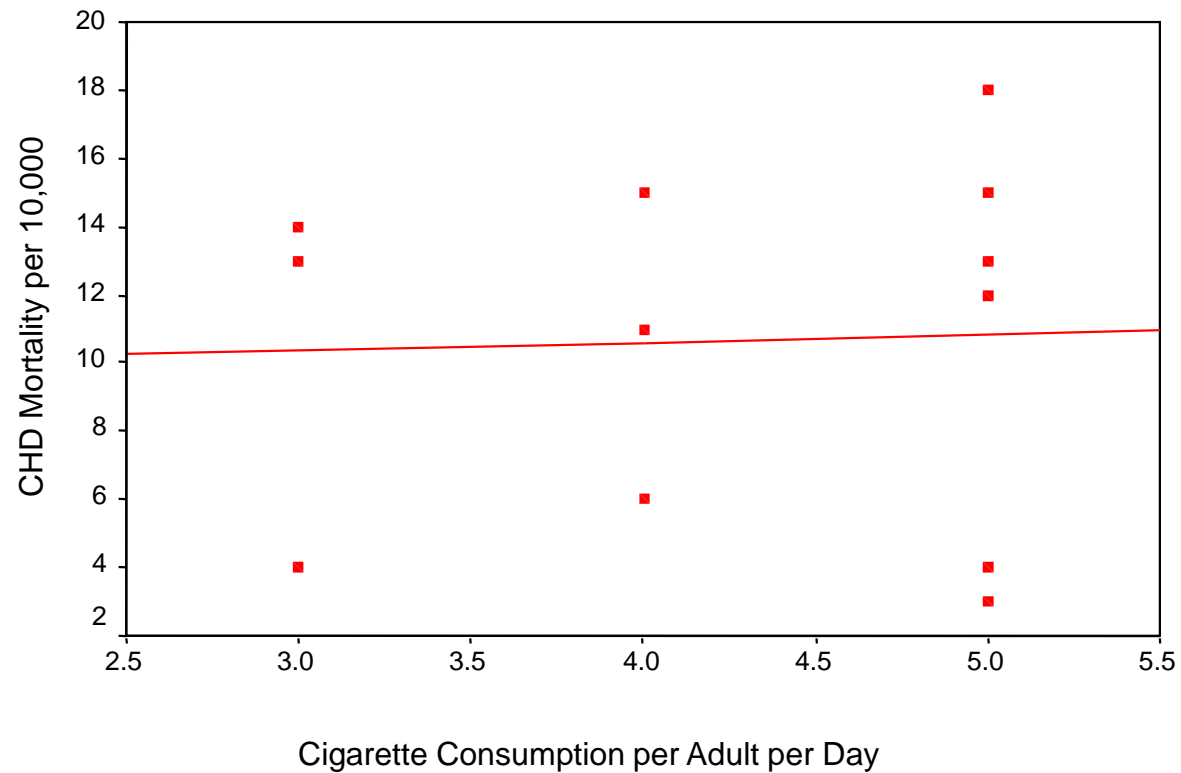
Truncation



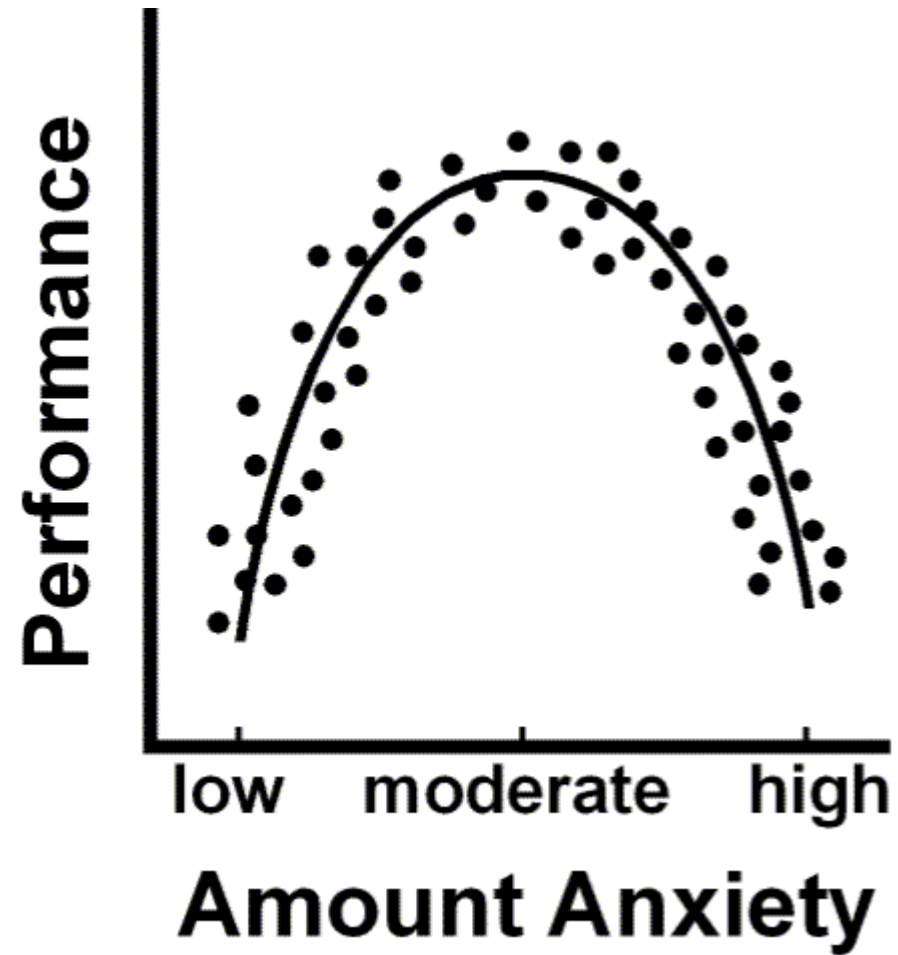
Countries With Low Consumptions

Data With Restricted Range

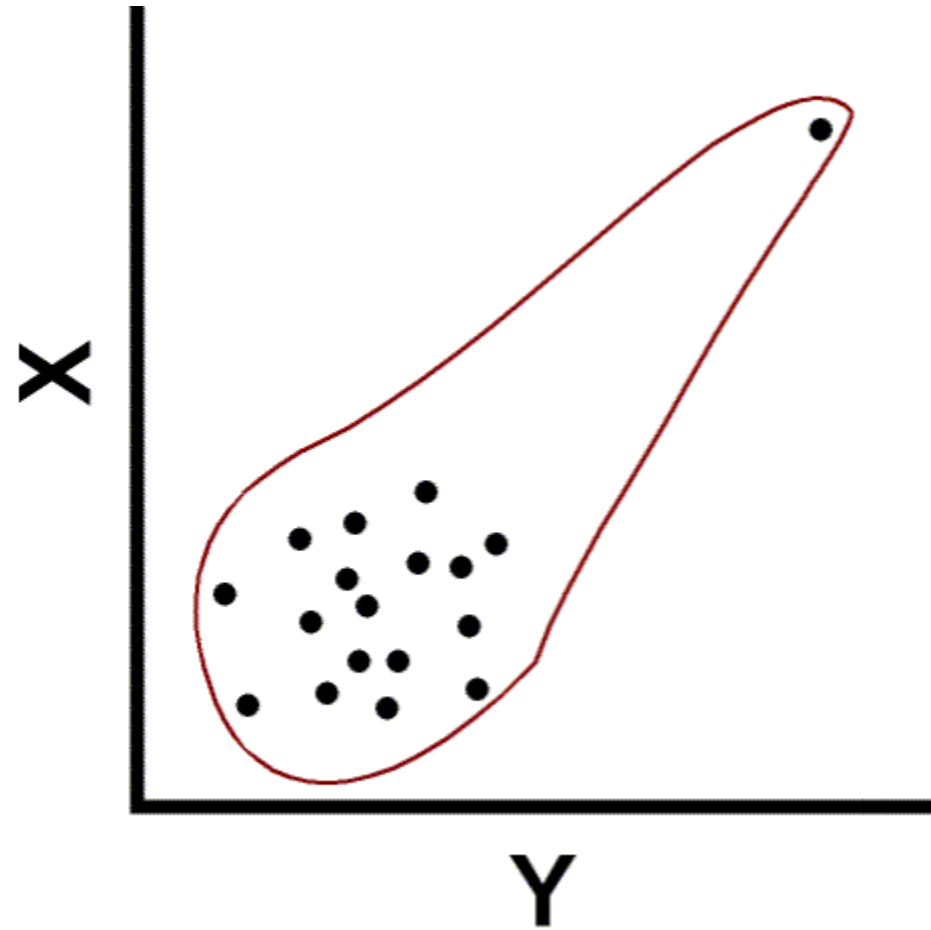
Truncated at 5 Cigarettes Per Day



Non-linearity



Outliers



Exercise (14)

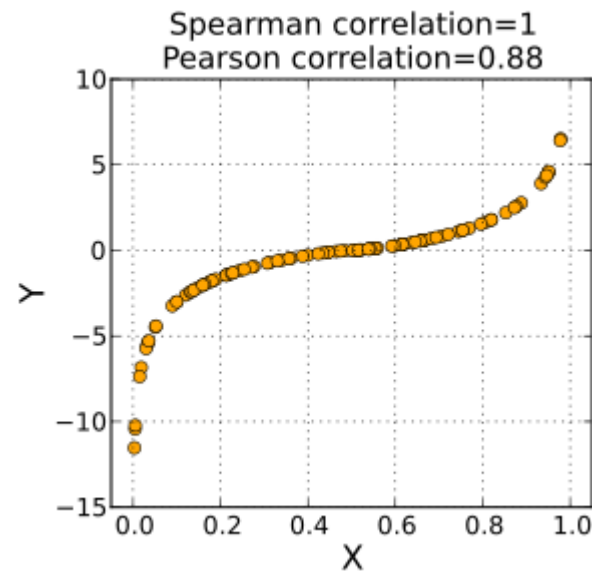
- ▶ Provide example data points matching the following description
- ▶ It should be constructed over $n=50$ data points in each variable
- ▶ Provide a table description of the example data as well as a scatter plot
- ▶ The $\text{Pearson}(x,y) > 0.9$ but where $n-1$ points can be selected so that for the vectors restricted to those we have Pearson correlation < 0.1

Correlation measures based on ranks

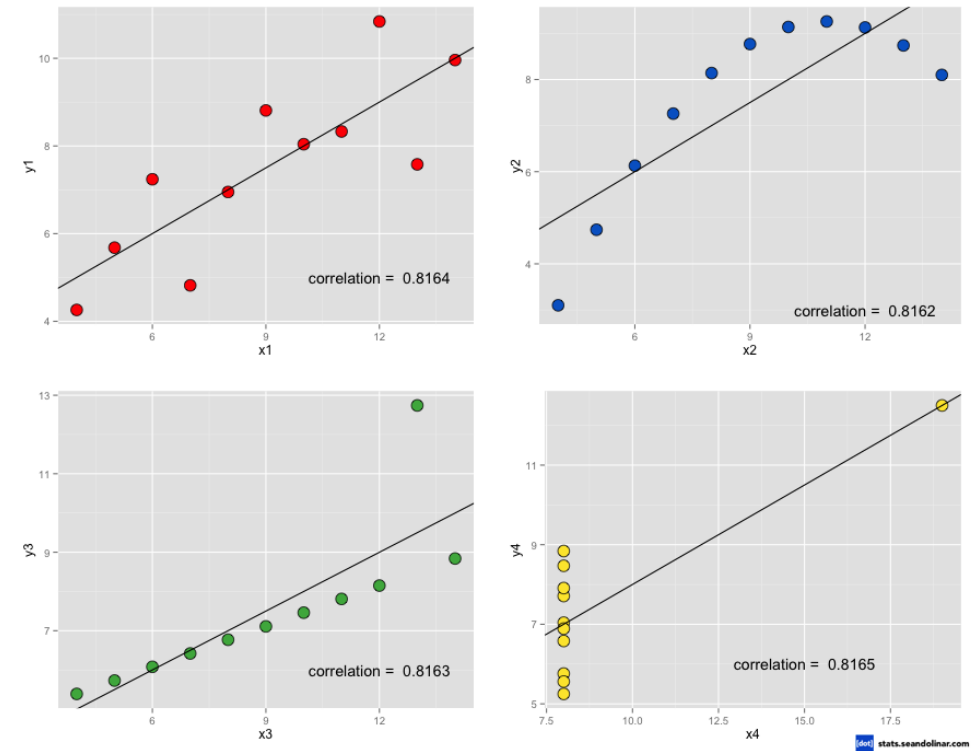
- ▶ Pearson correlation is very sensitive to the actual values of the data and to outliers
 - ▶ Pearson (and its relatives) measure linearity of the data. This is not always the correct model or desired observation/assessment (a strong non-linear relationship can exist but Pearson might not get too excited ...)
 - ▶ Statistical assessment thus heavily depends on the assumed null distributions
 - ▶ It is often much more robust to use rank based correlation measures
 - ▶ First step – transform the data into ranks
 - ▶ Next step – calculate summary statistic on the permutation obtained
 - ▶ Note – we will review several approaches and ties are handled in different ways for each of them
 - ▶ Last step – statistical significance (over the null $\text{Unif}(S_n)$)
- | | |
|-------------|----|
| 0.025528291 | 10 |
| 0.358797703 | 5 |
| 0.139868904 | 8 |
| 0.351813239 | 6 |
| 0.714498127 | 1 |
| 0.298464807 | 7 |
| 0.613028097 | 3 |
| 0.129888803 | 9 |
| 0.524701958 | 4 |
| 0.688437492 | 2 |

Spearman

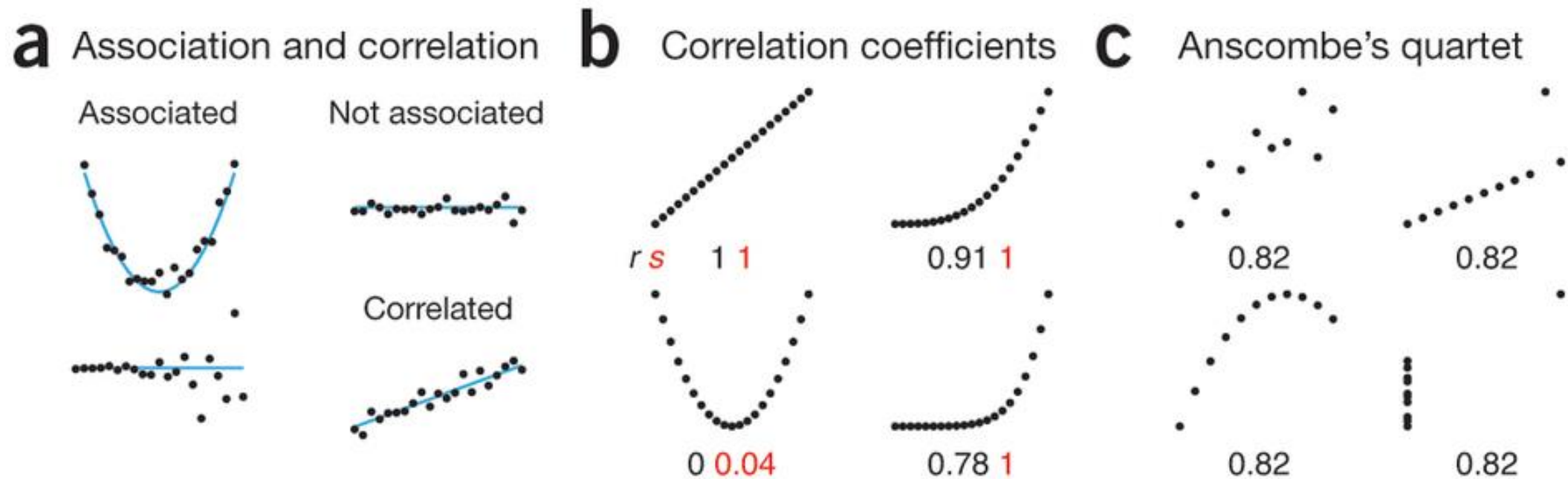
- ▶ Perform Pearson correlation on the rank values
- ▶ Ties can be handled by fractional ranks
- ▶ $-1 \leq \text{SRC} \leq 1$, always ...
- ▶ When is it -1? 1?



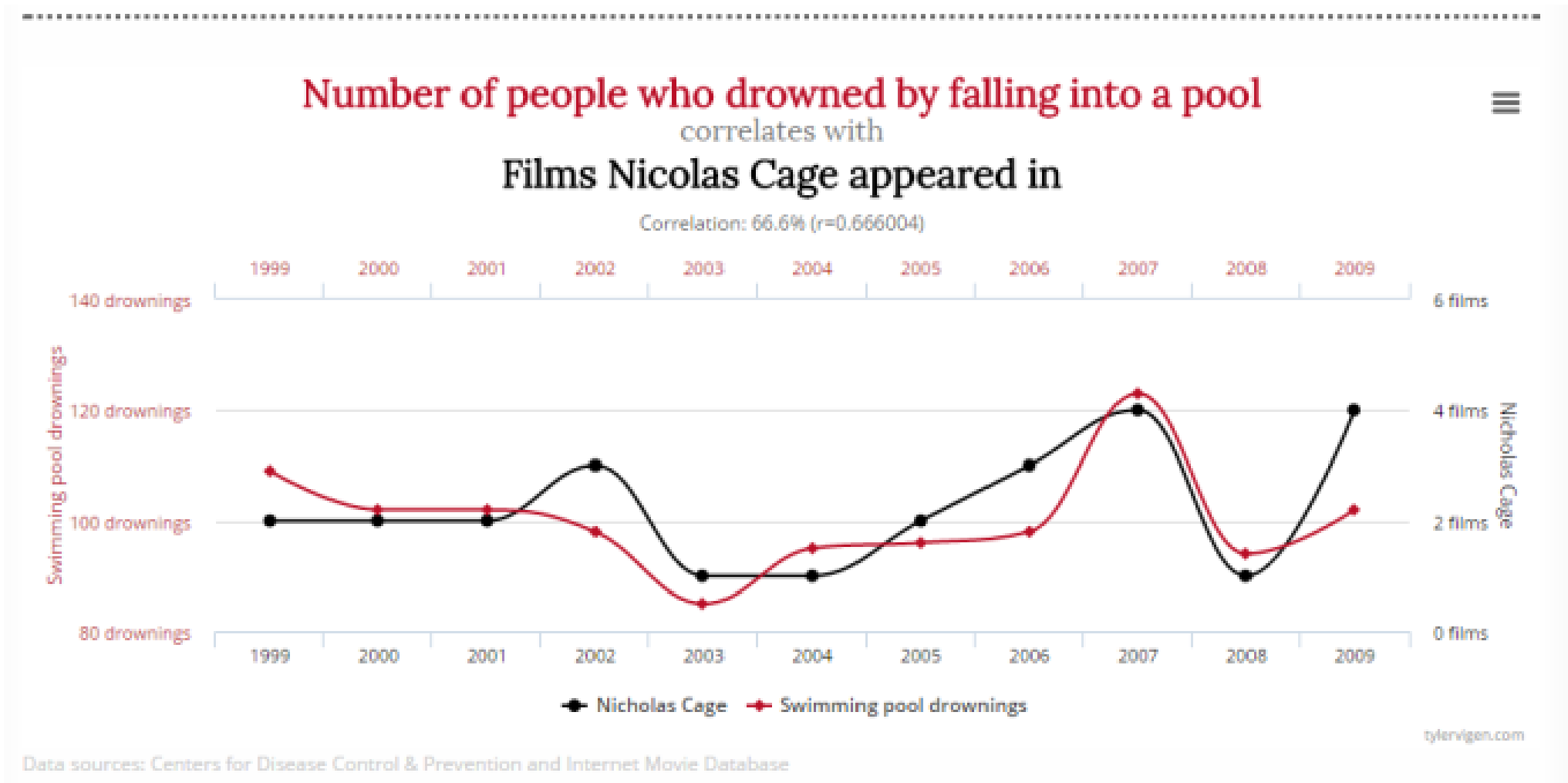
Anscombe Quadrant -- Correlation Demonstration



Modern association measures



Correlation is NOT causation



Exercise (15)

- ▶ Calculate Spearman on the same variables from Ex. 13

Exercise (16)

- ▶ Given the following dataset:

Country	Alcohol from Wine(x_i)	Heart Disease Deaths(y_i)	Country	Alcohol from Wine(x_i)	Heart Disease Deaths(y_i)
Australia	2.5	210	Netherlands	1.8	167
Austria	3.9	166	New Zealand	1.9	266
Belgium	2.9	131	Norway	0.8	227
Canada	2.4	191	Spain	6.5	86
Denmark	3.0	220	Sweden	1.6	207
Finland	0.85	297	Switzerland	5.8	115
France	9.1	71	U.K.	1.3	285
Iceland	0.75	211	U.S.	1.2	199
Ireland	0.7	300	W. Germany	2.7	172
Italy	7.9	107			

Exercise (16)

- ▶ Create a scatter plot of CVD mortality Vs Wine consumption
- ▶ Convert the values to ranks – use average ranks for ties
- ▶ Use the following formula to calculate the Spearman correlation:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

* Where $d_i = \text{rank}(x_i) - \text{rank}(y_i)$

- ▶ Calculate Spearman using SciPy

Kendall correlation coefficient

- ▶ Let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be the observed data
- ▶ Assume that all values of (x_i) and (y_i) are unique
- ▶ A pair of observations (x_i, y_i) and (x_j, y_j) is said to be *concordant* if the ranks for both elements agree: that is, if both $x_i > x_j$ and $y_i > y_j$ OR if both $x_i < x_j$ and $y_i < y_j$
- ▶ It is said to be *discordant*, if $x_i > x_j$ and $y_i < y_j$ OR if $x_i < x_j$ and $y_i > y_j$
- ▶ If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant. (and we assumed, for simplicity, that this doesn't happen)
- ▶ Let C and D be the number of concordant and discordant pairs, respectively
- ▶ The Kendall τ coefficient is defined as:

$$\tau = \frac{C - D}{\binom{n}{2}}$$

How to deal with ties?

$$\tau = \frac{C - D}{\sqrt{\left(\binom{n}{2} - t_x\right) \left(\binom{n}{2} - t_y\right)}}$$

- ▶ Where t_x and t_y are the number of ties in each of the dimensions.

Kendall's τ - example

- Grades of 11 students in 2 exams:

Exam 1	Exam 2
85	85
98	95
90	80
83	75
57	70
63	65
77	73
99	93
80	79
96	88
69	74

Ranks of exam results and calculating C and D

Exam1 x	Exam2 y	c	d
1	2	9	1
2	1	9	0
3	3	8	0
4	5	6	1
5	4	6	0
6	7	4	1
7	6	4	0
8	9	2	1
9	8	2	0
10	11	0	1
11	10	C=50	D=5

$$\tau = 0.818$$

Exercise (17)

- Use the following procedure to calculate Kendell on the previous data:

```

numer := 0
for i:=1..N-1 do
    for j:=i+1..N do
        numer := numer +
            sign(x[i] - x[j])
            * sign(y[i] - y[j])
return numer

```

Exam 1	Exam 2
85	85
98	95
90	80
83	75
57	70
63	65
77	73
99	93
80	79
96	88
69	74

Strength AND Significance

- ▶ We introduced mathematical measures of the **STRENGTH** of a relationship between two variables
- ▶ We **DIDN'T** discussed assessing statistical **SIGNIFICANCE**
- ▶ Note that a relationship can be **strong** and yet **not** significant
- ▶ Conversely, a relationship can be **weak** but **significant**
- ▶ The key factor is the **size of the sample**
- ▶ For small samples, it is easy to produce a strong correlation by chance and one must pay attention to significance to avoid jumping to spurious conclusions
- ▶ For large samples, it is easy to achieve significance, and one must pay attention to the strength of the correlation to determine if the relationship explains much about the data

Parameter Estimation - Motivation

- ▶ Given the following Dataset – 100 points in each class
- ▶ What will be the prediction according to Bayesian classifier (using counting)?
- ▶ Small recap:

- ▶ Bayes classifier uses MAP:

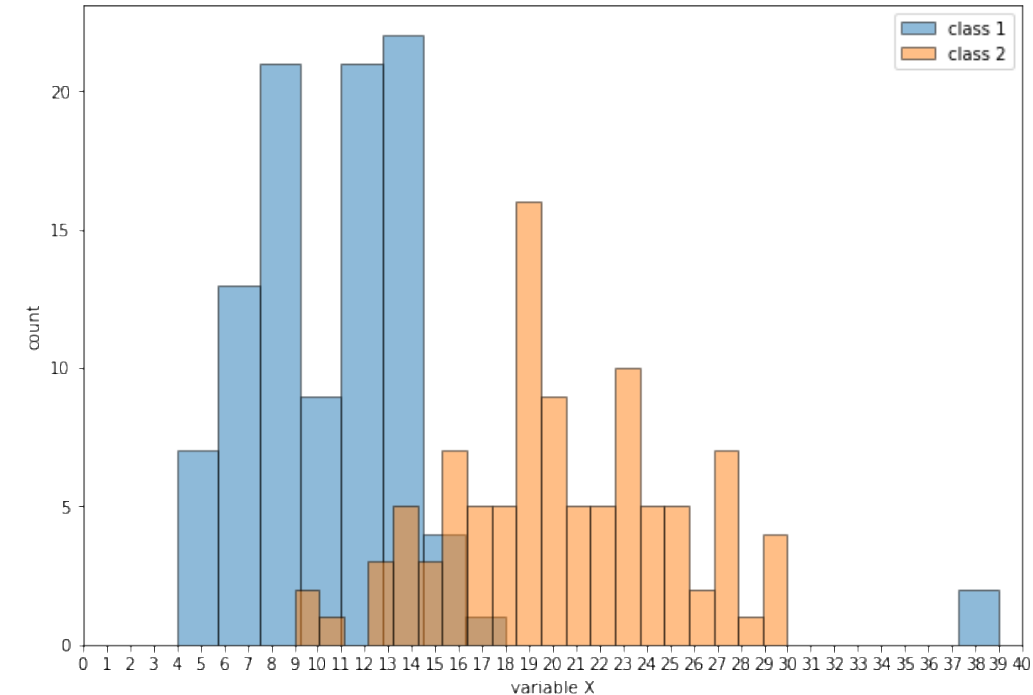
$$\operatorname{argmax}(P(A|x), P(B|x))$$

$$\operatorname{argmax}(P(x|A) * P(A), P(x|B) * P(B))$$

* The denominator is the same...

- ▶ Assuming the same Prior we get:

$$\operatorname{argmax}(P(x|A), P(x|B))$$



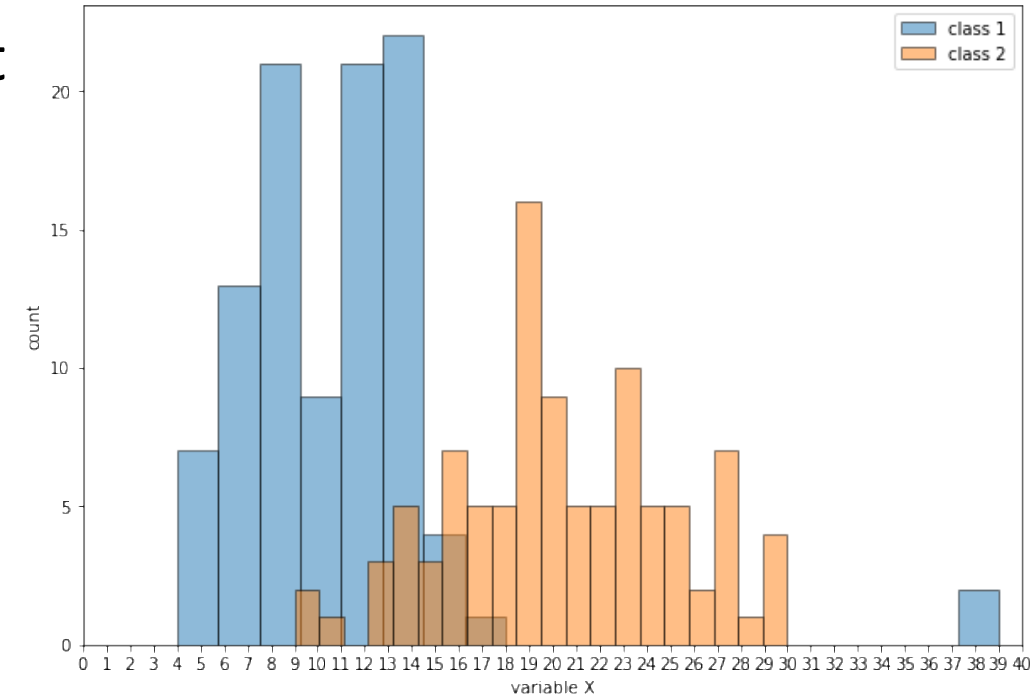
Motivation

- ▶ What will be the prediction of new data point where $x=9$?

$$P(x|A) = \frac{21}{100} = 0.21, P(x|B) = \frac{2}{100} = 0.02$$

- ▶ The prediction will be class A

* We need to use Laplace, but it won't change the prediction in this case

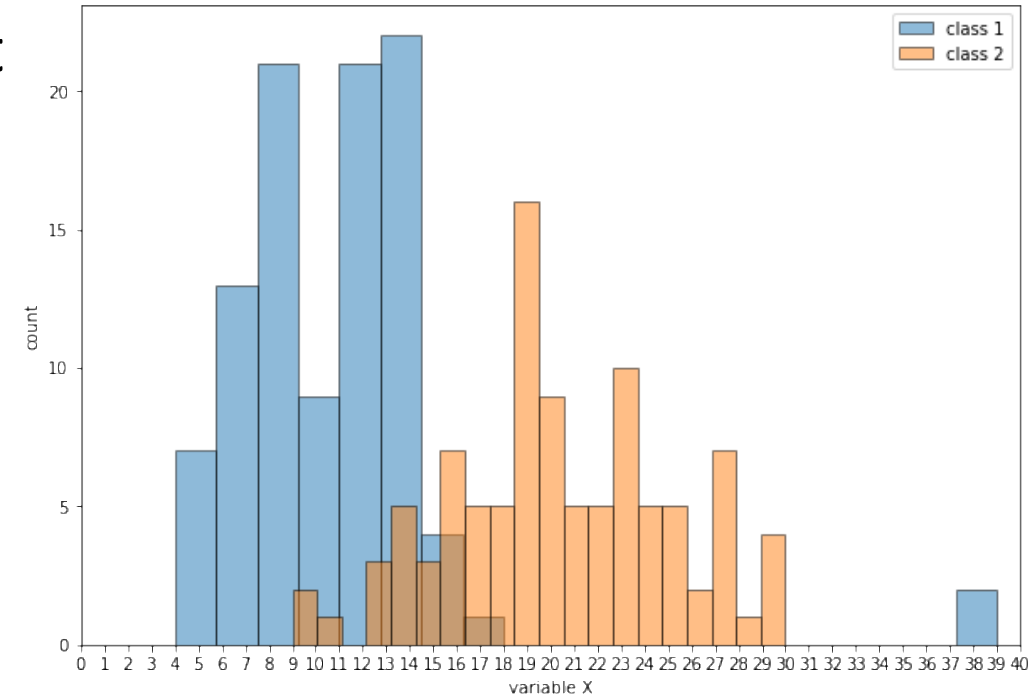


Motivation

- ▶ What will be the prediction of new data point where $x=38$?

$$P(x|A) = \frac{2}{100} = 0.02, P(x|B) = \frac{0}{100} = 0$$

- ▶ The prediction will be class A
- ▶ It seems like the generalization is not so good...
- ▶ We want a method that output a model with better generalization
- ▶ We will do it by find the underline distributions

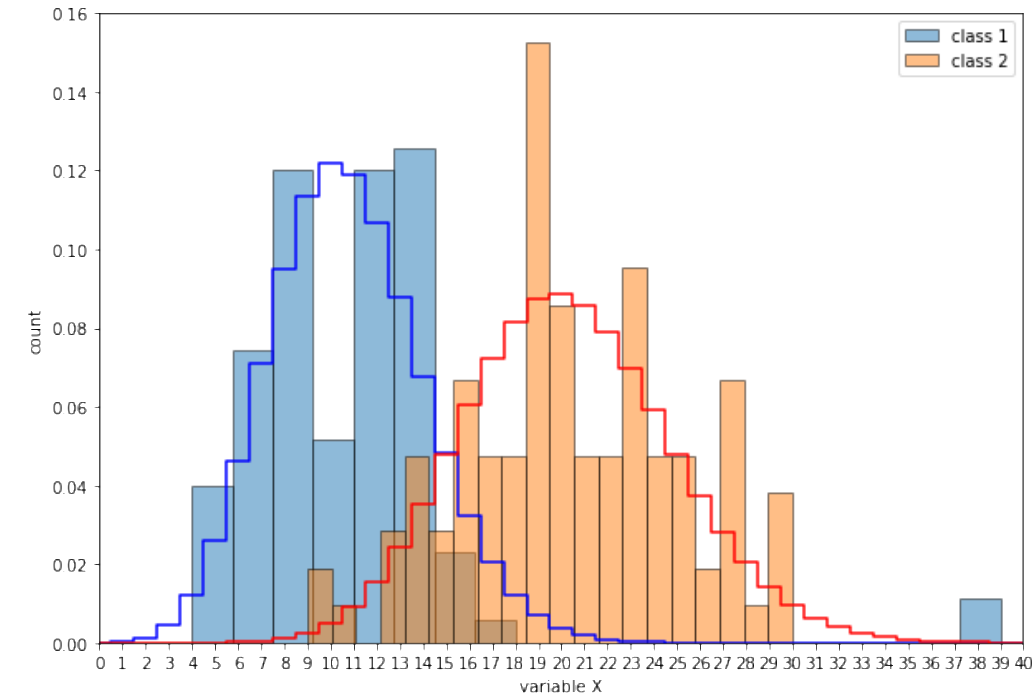


Motivation

- ▶ What we need to do?
 - ▶ Select distribution
 - ▶ Today visualization (in the future more principle approaches will be discussed)
 - ▶ Select distribution parameters
 - ▶ Poisson – λ
 - ▶ Which λ is better for Class A? 8, 10.5 or 11.7?
 - ▶ How we will decide?
- ▶ MLE
 - ▶ In this case the MLE finds the following λ :
 - Class A – 10.74
 - Class B – 20.31
 - ▶ What will be the prediction of new data point where $x=38$ according to MLE?

$$P(x|A) = \text{Poisson}(10.74).pmf(38) = 6.24 * 10^{-11}$$

$$P(x|B) = \text{Poisson}(20.31).pmf(38) = 1.42 * 10^{-4}$$
 - ▶ The prediction will be class B



MLE

- ▶ A straightforward approach to parameter estimation
- ▶ Directly works in simple cases
- ▶ Forms the basis for most parameter estimation approaches

- ▶ Given
 - ▶ A set of observed values $D = \{x_1, \dots, x_n\}$
 - ▶ A model and a vector of parameters for this model, θ

- ▶ We define
 - ▶ The likelihood of the model given the data, $L(\theta | D)$, which we define to be $P(D | \theta)$ (note – no prior assumption)
 - ▶ Log-likelihood of the model given the data is often more useful and we write:
 $L(\theta) = \log P(D | \theta)$

- ▶ In MLE we seek $\theta_{ML} = \arg \max_{\theta \in \Omega} L(\theta)$

MLE for independent instances

- ▶ We often assume that the data instances are a result of independent identically distributed (i.i.d.) random variables. This is the same as assuming independent repeats of the same generation mechanism

$$\begin{aligned}\theta_{ML} &= \arg \max_{\theta \in \Omega} L(\Theta) \\ &= \arg \max_{\theta \in \Omega} \log P(D | \Theta) \\ &= \arg \max_{\theta \in \Omega} \log P(x_1, \dots, x_n | \Theta) \\ &= \arg \max_{\theta \in \Omega} \log \prod_i P(x_i | \Theta) \\ &= \arg \max_{\theta \in \Omega} \sum_i \log P(x_i | \Theta)\end{aligned}$$

- ▶ Solving the optimization problem varies in its complexity, depending on the form of $p(x|\theta)$ – e.g the (common) pdf of the underlying variables

First case

- ▶ Assuming
 - ▶ A coin has a probability p of being heads, $1-p$ of being tails
 - ▶ Observation:
 - ▶ We toss the coin until the first head
 - ▶ That is we are observing $K \sim \text{Geo}(p)$

- ▶ What is the value of p based on MLE, given that the first success was in the k -th toss?

Coin tossing

$$L(\Theta) = \log P(D|\Theta) = \log(1 - p)^{k-1} p^1$$

$$= (k - 1) \log(1 - p) + \log p$$

$$\frac{dL(\Theta)}{dp} = -\frac{k - 1}{1 - p} + \frac{1}{p}$$

$$\frac{dL(\Theta)}{dp} = \frac{-kp + p + 1 - p}{(1 - p)p} = 0$$

$$1 = kp$$

$$p = \frac{1}{k}$$

Coin tossing – N independent experiments

$$\begin{aligned} L(\Theta) &= \log P(D|\Theta) = \sum_{i=1}^N \log(1-p)^{k_i-1} p^1 \\ &= \sum_{i=1}^N ((k_i - 1) \log(1-p) + \log p) \\ &= \log(1-p) \left(\left(\sum_{i=1}^N k_i \right) - N \right) + N \log p \\ \frac{dL(\Theta)}{dp} &= -\frac{\sum_{i=1}^N k_i - N}{1-p} + \frac{N}{p} \\ \frac{dL(\Theta)}{dp} &= \frac{-p \sum_{i=1}^N k_i + Np + N - Np}{(1-p)p} = 0 \\ N &= p \sum_{i=1}^N k_i \rightarrow p = \frac{N}{\sum_{i=1}^N k_i} = \frac{1}{\frac{\sum_{i=1}^N k_i}{N}} = \frac{1}{\bar{K}} \end{aligned}$$

Next case

- ▶ Assuming
 - ▶ A coin has a probability p of being heads, $1-p$ of being tails
 - ▶ Observation:
 - ▶ We toss the coin N times, observing a set of Hs and Ts.
- ▶ What is the value of p based on MLE, given the observation?

$$\begin{aligned}L(\Theta) &= \log P(D \mid \Theta) = \log p^m (1-p)^{N-m} \\ &= m \log p + (N-m) \log(1-p)\end{aligned}$$

$$\frac{dL(\Theta)}{dp} = \frac{d(m \log p + (N-m) \log(1-p))}{dp} = \frac{m}{p} - \frac{N-m}{1-p} = 0$$

$$p = m/N$$

- ▶ What is missing in the equation?

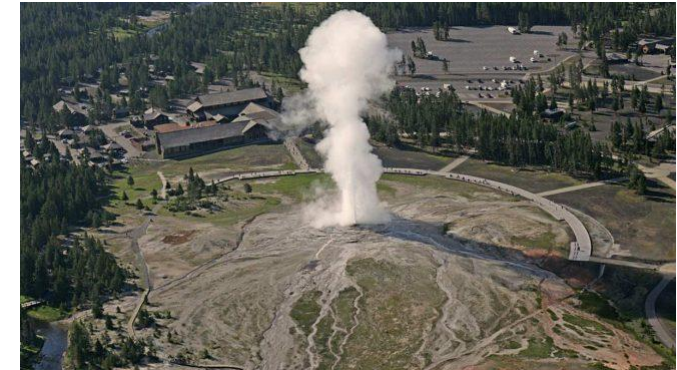
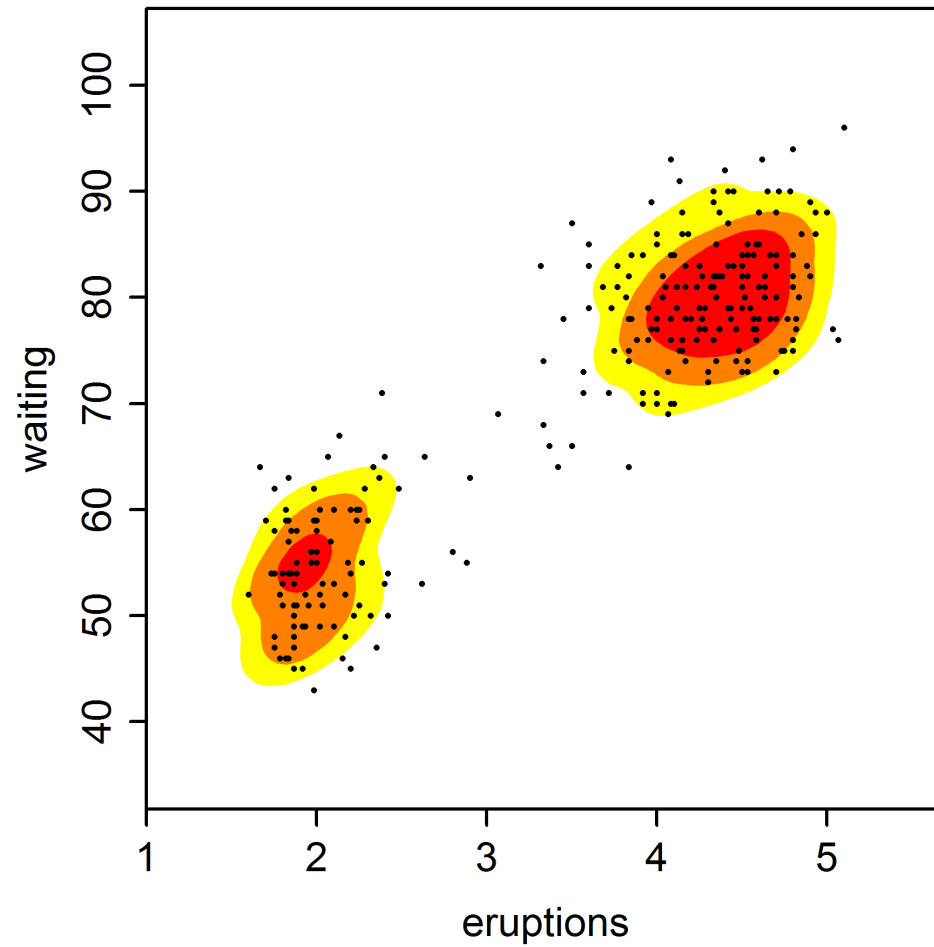
Another simple example: Poisson

$$\begin{aligned} L(\Theta) &= \log P(D|\Theta) \\ &= \log \prod_{j=1}^n e^{-\lambda} \frac{\lambda^{x_j}}{x_j!} \\ &= \sum_{j=1}^n \log e^{-\lambda} \frac{\lambda^{x_j}}{x_j!} \\ &= \sum_{j=1}^n (\log e^{-\lambda} + \log \lambda^{x_j} - \log x_j!) \\ &= \sum_{j=1}^n (-\lambda + x_j \log \lambda - \log x_j!) = -n\lambda + \log \lambda \sum_{j=1}^n x_j - \sum_{j=1}^n \log x_j! \\ \frac{dL(\Theta)}{d\lambda} &= -n + \frac{1}{\lambda} \sum_{j=1}^n x_j = 0 \\ \frac{1}{n} \sum_{j=1}^n x_j &= \lambda \end{aligned}$$

Expectation Maximization (EM) Algorithm

- ▶ Iterative method for parameter estimation where layers of data are missing from the observation
- ▶ Dempster, Laird, Rubin, J of the Royal Stat Soc, 1977
- ▶ Many variations followed. Research into methodology and applications is very active
- ▶ Has two steps:
 - ▶ Expectation (E) and Maximization (M)
- ▶ Applicable to a wide range of machine learning and inference tasks

Example

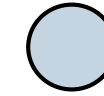


Basic setting in EM

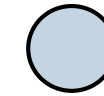
- ▶ D is a set of data points: **observed** data
- ▶ Θ is a parameter vector.
- ▶ EM is an iterative method for finding θ_{ML}
- ▶ It's mostly useful when
 - ▶ Calculating $P(x \mid \theta)$ directly is hard.
 - ▶ Calculating $P(x, z \mid \theta)$ is simpler, where z is some “hidden” data (or “missing” data)
 - ▶ The hidden data is assumed to be determined by some other rv Z , which is part of the model.
 - ▶ Note: the model, under θ , controls both X and Z but in D we only see the values of X .

Randomly selecting one of two coins

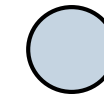
- ▶ There are two coins, with p_A and p_B
- ▶ One of the coins is selected, with w_A and w_B probabilities
- ▶ Then it is tossed 10 times
- ▶ We observe the results of many repeats of this exercise
- ▶ If we know which coin is tossed in each set then we can do MLE and get both the p s and the w s
- ▶ But we don't ...
- ▶ Lets see what EM can do here



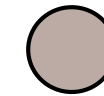
HHHHTHHHHH



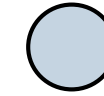
THHHHHHHHTH



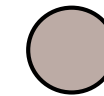
HHHHHHHHTHH



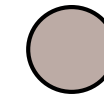
HHTHTTHHTT



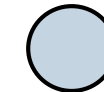
HHTHHHHHHTH



HTTHTHHHTT



HTTHTHHHHT



HTHHTHHHHT

The EM algorithm

- ▶ Consider a set of starting parameters, including the parameters of Z
- ▶ Use these to “estimate” the values of the missing data, per observed data point
- ▶ Use the “complete” data to update all parameters (of both Z and $X|Z$)
- ▶ Repeat until convergence

EM: uncovering the coins ...

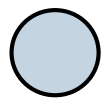
$$P_A(x_1) = w_A \binom{10}{9} 0.6^9 0.4^1 = 0.04$$

$$P_B(x_1) = w_B \binom{10}{9} 0.5^9 0.5^1 = 0.01$$

$$r(x_1, A) = \frac{0.04}{0.05} = 0.8$$

$$r(x_1, B) = \frac{0.01}{0.05} = 0.2$$

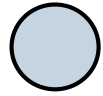
Note: use aposteriori estimates



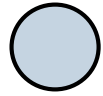
HHHHTHHHHH

0.8

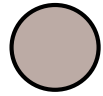
0.2



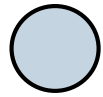
THHHHHHHHTH



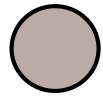
HHHHHHHHTHH



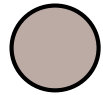
HHTHTTTHHTT



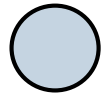
HHTHHHHHHTH



HTTHTHHHTT

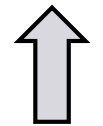


HTTHTHHHHHT



HTHHHTHHHHHT

1



Init $p_A = 0.6$
 $p_B = 0.5$
ws are 0.5

2

Compute responsibilities

Coin A responsibilities

Coin B responsibilities

EM: uncovering the coins ...

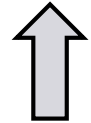
$$P_A(x_2) = w_A \binom{10}{8} 0.6^8 0.4^2 = 0.12$$

$$P_B(x_2) = w_B \binom{10}{8} 0.5^8 0.5^2 = 0.044$$

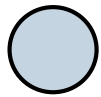
$$r(x_2, A) = \frac{0.12}{0.164} = 0.76$$

$$r(x_2, B) = \frac{0.044}{0.164} = 0.24$$

I



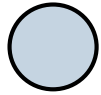
Init $p_A = 0.6$
 $p_B = 0.5$
ws are 0.5



HHHHTHHHHH

0.8

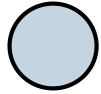
0.2



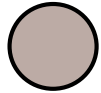
THHHHHHHHTH

0.76

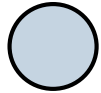
0.24



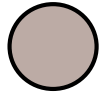
HHHHHHHTHH



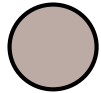
HHTHTTHTTT



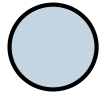
HHTHHHHHTH



HTTHTHHHTT



HTTHTHHHHT

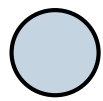


HTHHHTHHHHT

Coin A responsibilities

Coin B responsibilities

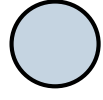
EM: uncovering the coins ...



HHHHTHHHHH

0.8

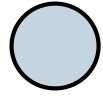
0.2



THHHHHHHHTH

0.76

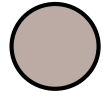
0.24



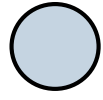
HHHHHHHHTHH

0.8

0.2



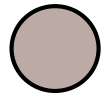
HHTHTTHTTT



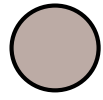
HHTHHHHHHTH

0.76

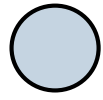
0.24



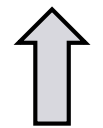
HTTHTHHHTT



HTTHTHHHHT



HTHHHTHHHHT



Coin A responsibilities

Coin B responsibilities

Init $p_A = 0.6$
 $p_B = 0.5$
ws are 0.5

EM: uncovering the coins ...

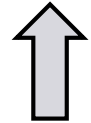
$$P_A(x_4) = w_A \binom{10}{5} 0.6^5 0.4^5$$

$$P_B(x_4) = w_B \binom{10}{5} 0.5^5 0.5^5$$

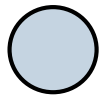
$$r(x_4, A) = 0.45$$

$$r(x_4, B) = 0.55$$

I



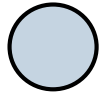
Init $p_A = 0.6$
 $p_B = 0.5$
ws are 0.5



HHHHTHHHHH

0.8

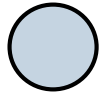
0.2



THHHHHHHHTH

0.76

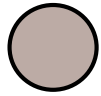
0.24



HHHHHHHHTHH

0.8

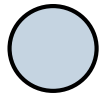
0.2



HHTHTTHTTT

0.45

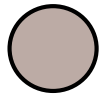
0.55



HHTHHHHHHTH

0.76

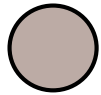
0.24



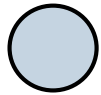
HTTHTHHHTT

0.45

0.55



HTTHTHHHHT



HTHHHTHHHHT

Coin A responsibilities

Coin B responsibilities

EM: uncovering the coins ...

3

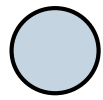
Compute new assignments:

$$\text{New } w_A = \frac{1}{N} \sum_{i=1}^N r(x_i, A)$$

$$\text{New } w_B = \frac{1}{N} \sum_{i=1}^N r(x_i, B)$$

$$\text{New } w_A = \frac{1}{8} \sum_{i=1}^8 r(x_i, A) = \frac{5.2}{8} = 0.65$$

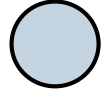
$$\text{New } w_B = \frac{1}{8} \sum_{i=1}^8 r(x_i, B) = \frac{2.8}{8} = 0.35$$



HHHHTHHHHH

0.8

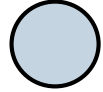
0.2



THHHHHHHHTH

0.76

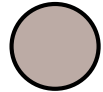
0.24



HHHHHHHTHH

0.8

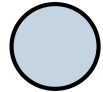
0.2



HHTHTTHTTT

0.45

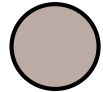
0.55



HHTHHHHHTH

0.76

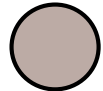
0.24



HTTHTHHHTT

0.45

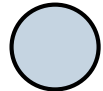
0.55



HTTHTHHHHT

0.55

0.45

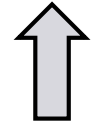


HTHHHTHHHHT

0.64

0.36

1



Init $p_A = 0.6$
 $p_B = 0.5$
ws are 0.5

2

Compute responsibilities

Coin A responsibilities

Coin B responsibilities

EM: uncovering the coins ...









3+4 Compute MLEs for the model parameters:

$$p_A = \frac{1}{(\text{New } w_A)N} \sum_{i=1}^N r(x_i, A)v(i)$$

$$p_B = \frac{1}{(\text{New } w_B)N} \sum_{i=1}^N r(x_i, B)v(i)$$

$$p_A = \frac{1}{5.2} \sum_{i=1}^8 r(x_i, A)v(i) = 0.745$$

$$p_B = \frac{1}{2.8} \sum_{i=1}^8 r(x_i, B)v(i) = 0.649$$

	HHHHTHHHHH	0.8	0.2	0.9
	THHHHHHHHTH	0.76	0.24	0.8
	HHHHHHHTHH	0.8	0.2	0.9
	HHTHTTHTTT	0.45	0.55	0.5
	HHTHHHHHHTH	0.76	0.24	0.8
	HTTHTHHHTT	0.45	0.55	0.5
	HTTHTHHHHHT	0.55	0.45	0.6
	HTHHHTHHHHHT	0.64	0.36	0.7

1 ↑

2

↑
 $r(x_i, A)$

↑
 $r(x_i, B)$

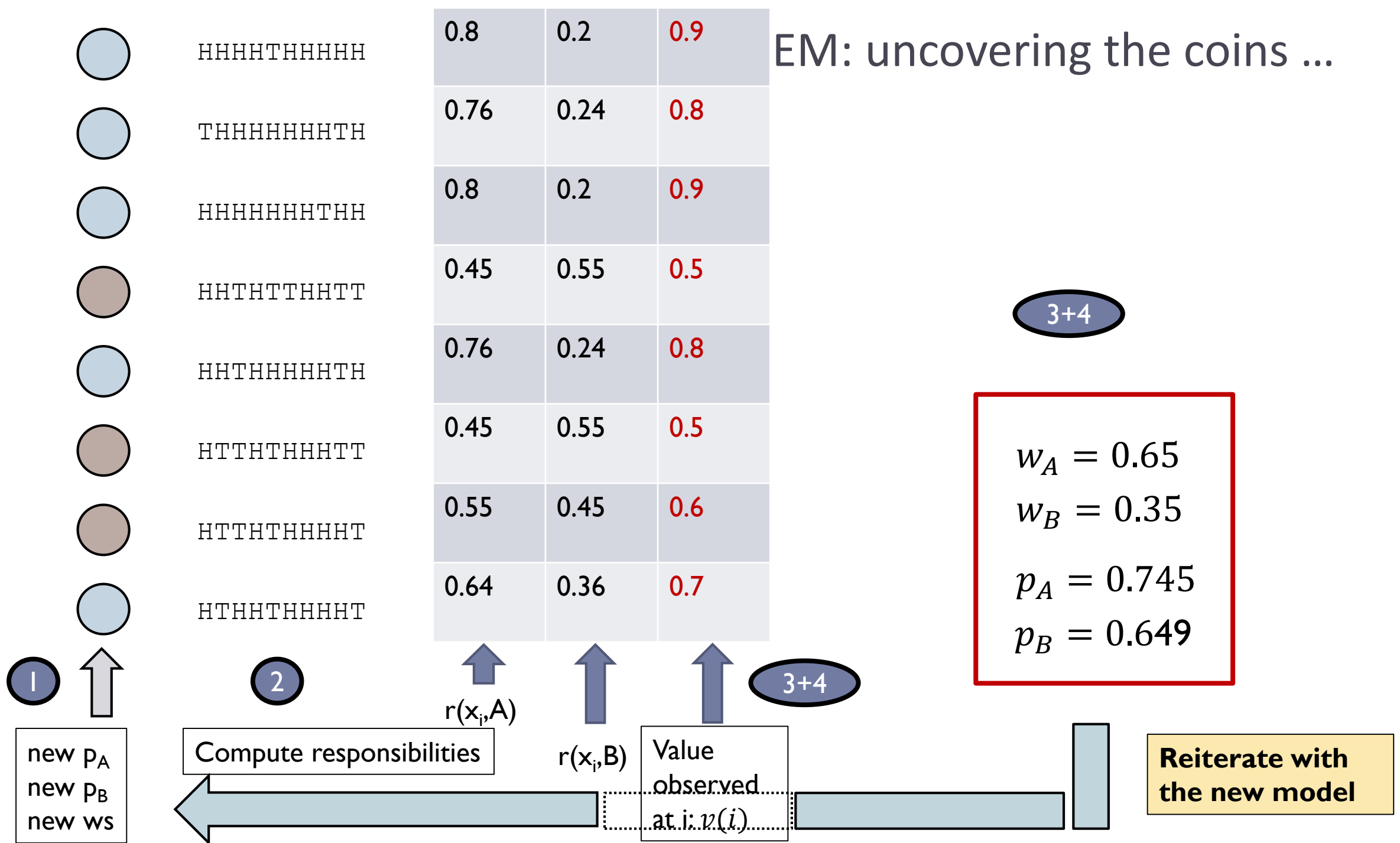
↑

3+4

Value
observed
at i : $v(i)$

Init $p_A = 0.6$
 $p_B = 0.5$
ws are 0.5

EM: uncovering the coins ...



The EM algorithm for two coins

- ▶ Consider a set of starting parameters, including the parameters of Z
- ▶ Use these to “estimate” the values of the missing data, per observed data point
 - ▶ Compute responsibilities using MAP
- ▶ Use the “complete” data to update all parameters (of both Z and X|Z)

$$\text{New } w_A = \frac{1}{N} \sum_{i=1}^N r(x_i, A)$$

$$p_A = \frac{1}{(\text{New } w_A)N} \sum_{i=1}^N r(x_i, A)v(i)$$

$$\text{New } w_B = \frac{1}{N} \sum_{i=1}^N r(x_i, B)$$

$$p_B = \frac{1}{(\text{New } w_B)N} \sum_{i=1}^N r(x_i, B)v(i)$$

- ▶ Repeat until convergence

EM for GMMs

- ▶ Step 1: Expectation (E-step)
 - ▶ Evaluate the “responsibilities” of each data point to each Gaussian using the current parameters
- ▶ Step 2: Maximization (M-step)
 - ▶ Re-estimate parameters (w_s , μ_s and σ_s) using the existing “responsibilities”
 - ▶ That is – every data point, x , contributes to each Gaussian component, G_i , in proportion to its responsibility: $r(x, G_i)$

Gaussian mixtures equations

- ▶ Responsibilities:

$$r(x, k) = \frac{w_k N(x | \mu_k, \sigma_k)}{\sum_{j=1}^K w_j N(x | \mu_j, \sigma_j)}$$

- ▶ Weights:

$$New\ w_j = \frac{1}{N} \sum_{i=1}^N r(x_i, j)$$

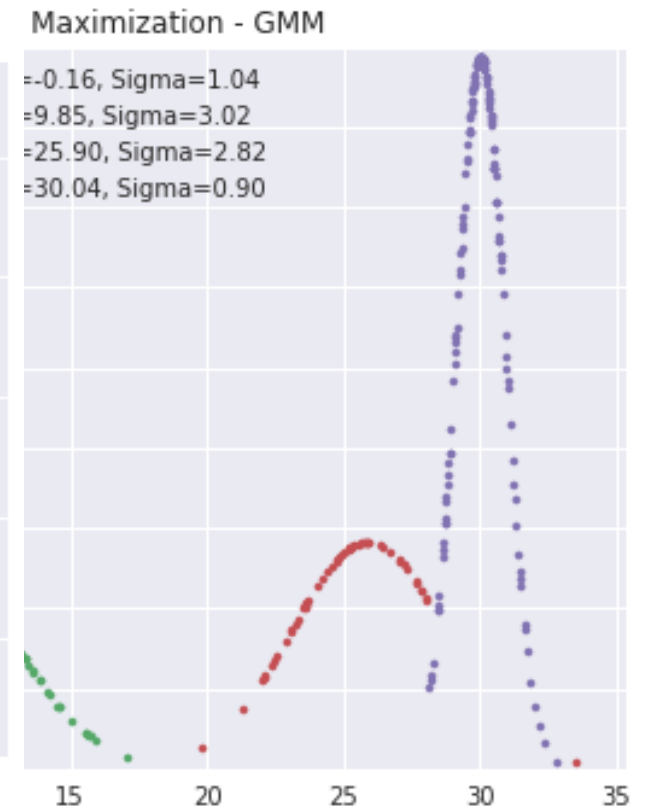
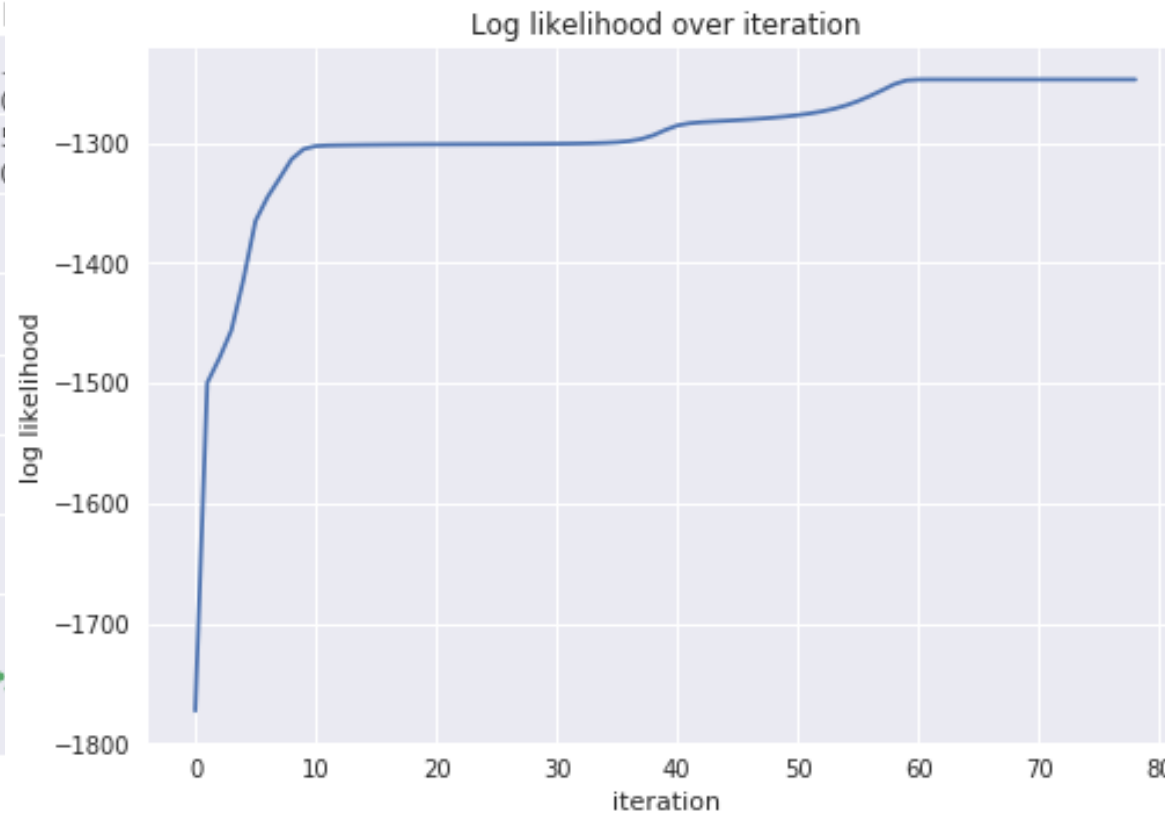
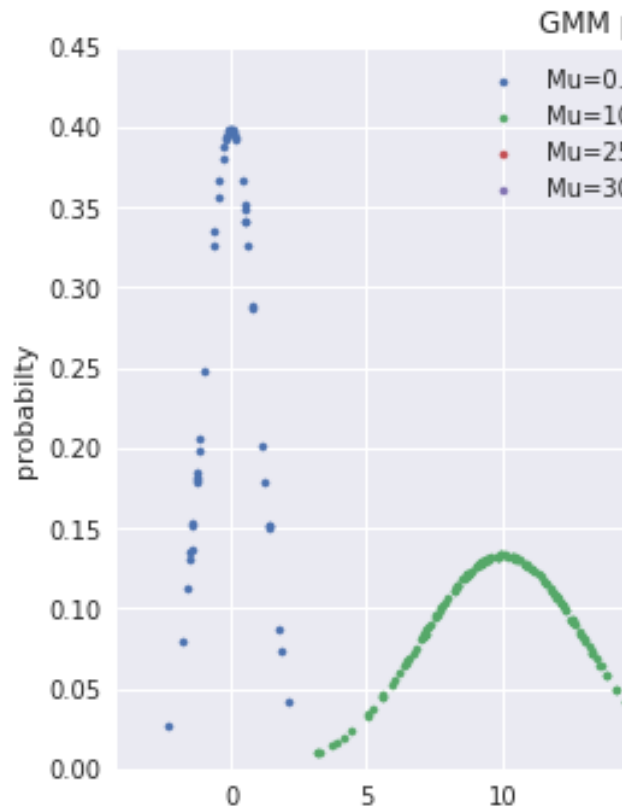
- ▶ Mean:

$$New\ \mu_j = \frac{1}{(New\ w_j)N} \sum_{i=1}^N r(x_i, j) x_i$$

- ▶ Variance:

$$(New\ \sigma_j)^2 = \frac{1}{(New\ w_j)N} \sum_{i=1}^N r(x_i, j) (x_i - New\ \mu_j)^2$$

Running example

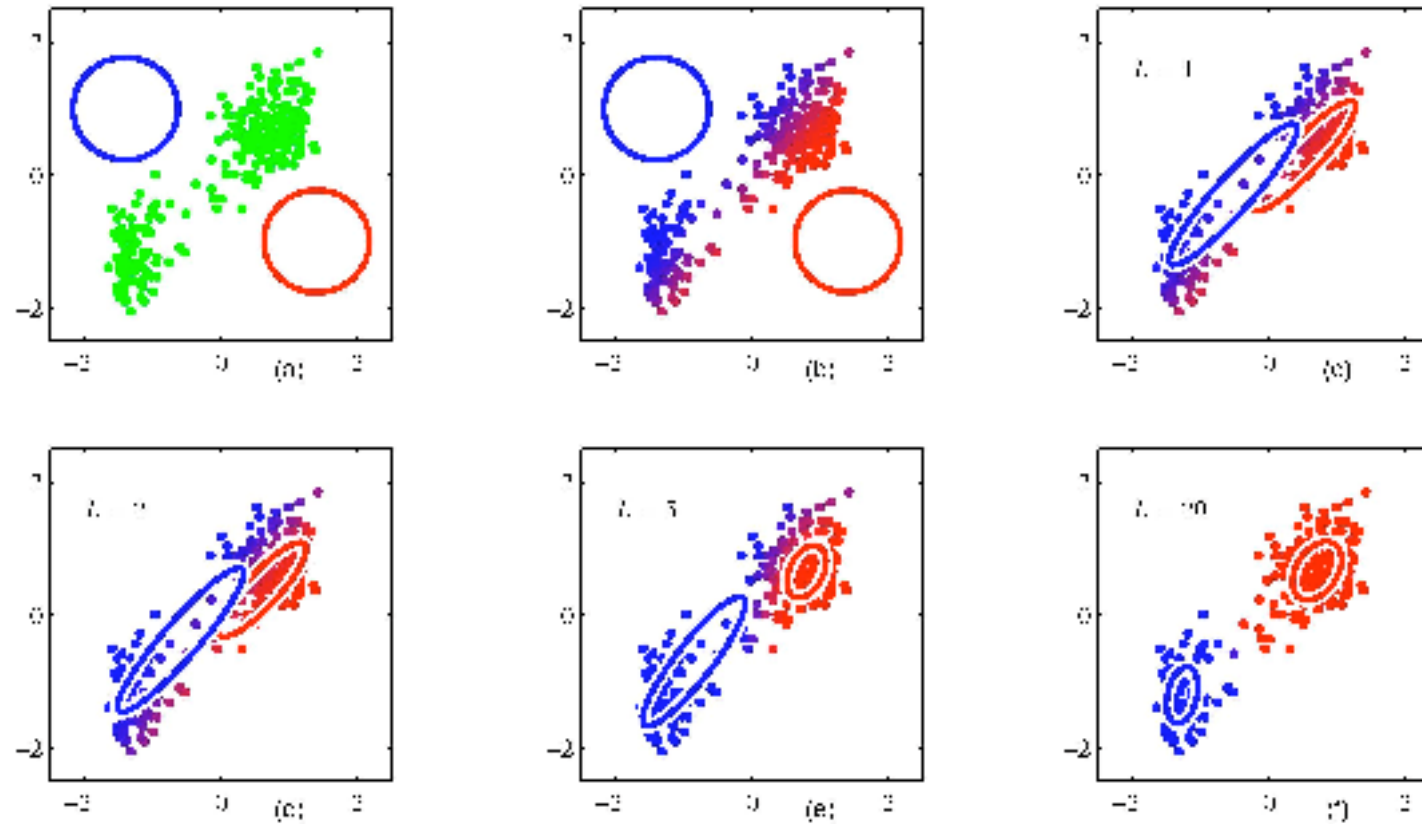


EM algorithm

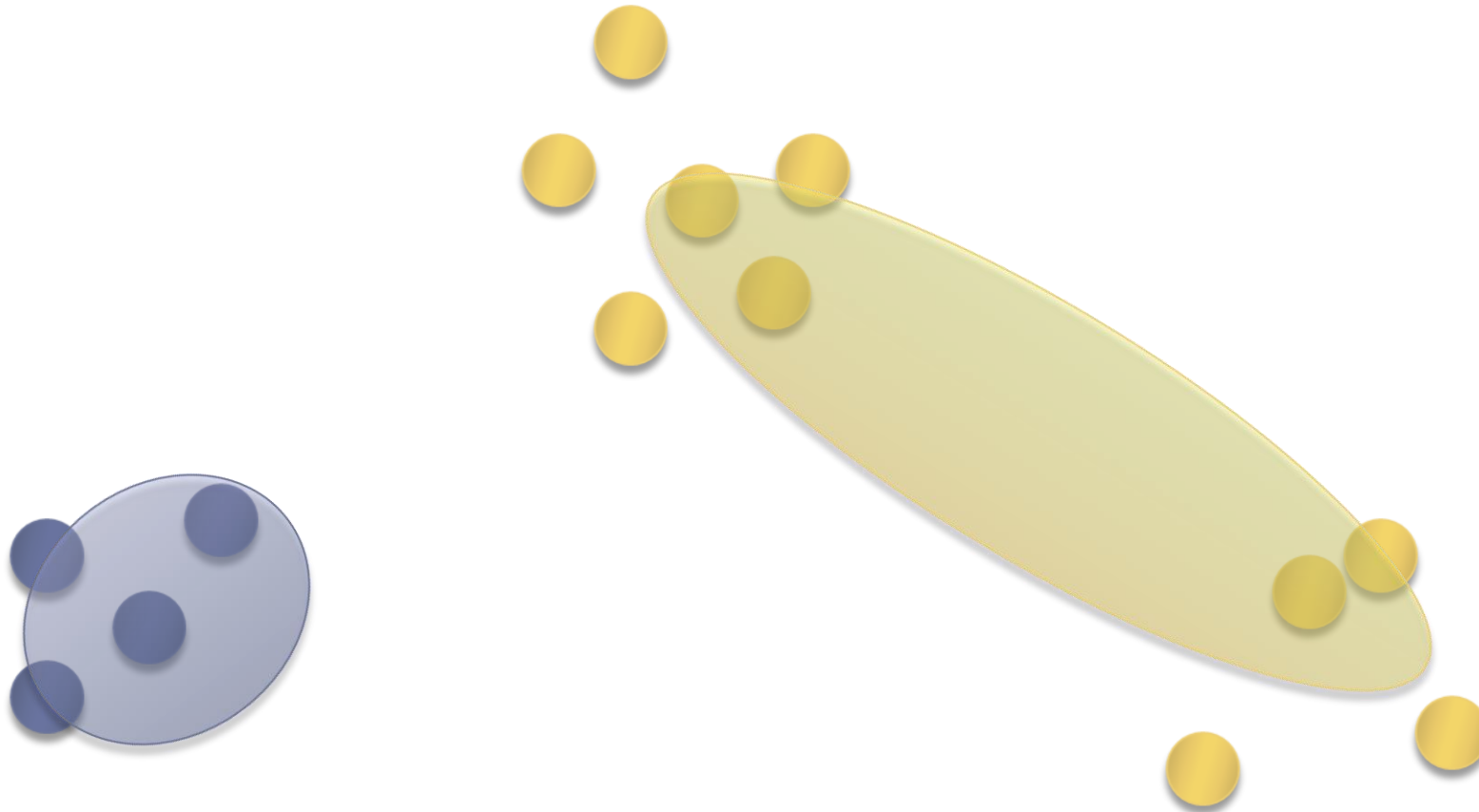
- ▶ A general algorithm/framework for inference where observations are dependent on a hidden intermediate
- ▶ Requires “specialization” to any given task or configuration

- Consider a set of starting parameters, including the parameters of Z
- Use these to compute responsibilities
- Use the “complete” data to update all parameters (of both Z and $X|Z$)
- Repeat until convergence

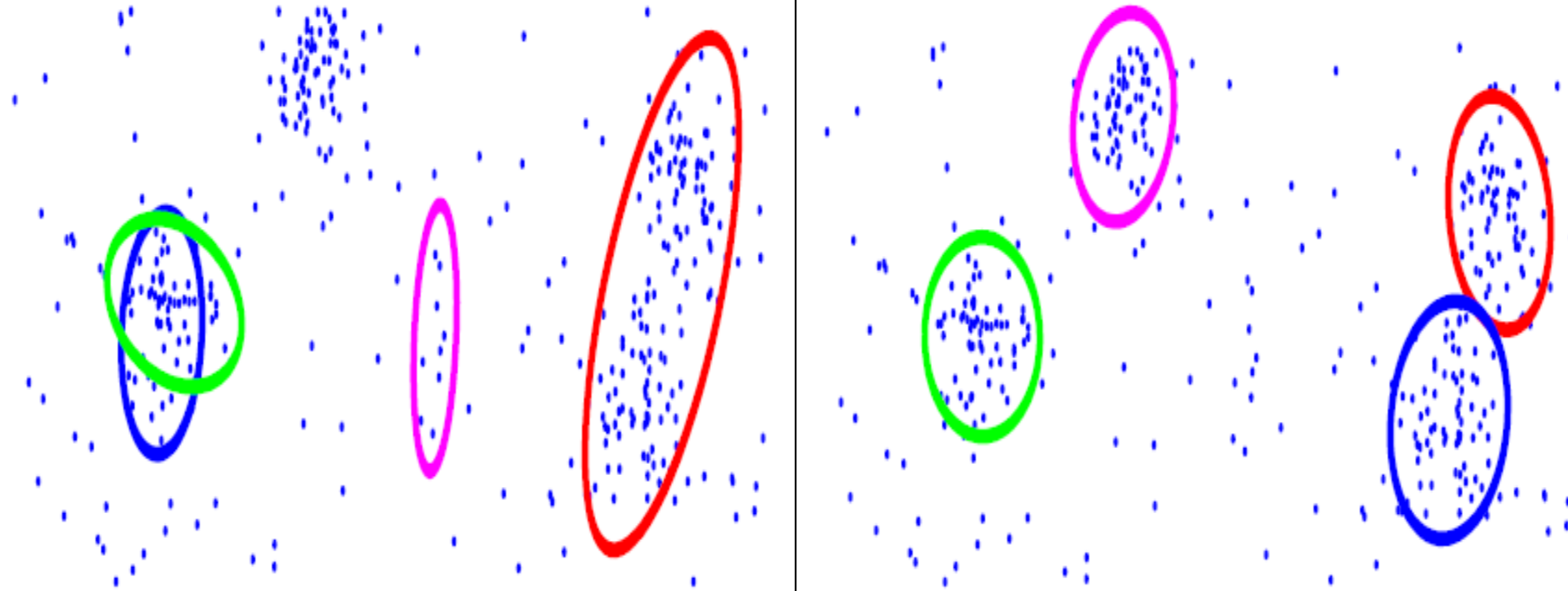
Visual example of EM



Incorrect Number of Gaussians



Local minima



EM application examples

- ▶ Clustering
- ▶ Modelling
- ▶ Prediction
- ▶ Outlier detection (predictive maintenance)



EM pros and cons

- ▶ Convergence: in every iteration of the EM algorithm the attained likelihood is improved over the previous round
- ▶ Is appropriate for (almost) any family of models and for any number of parameters
- ▶ Convergence can be very slow on some instances and is intimately related to the amount of missing information
- ▶ Like any learning approach, we work on the training data. It is important to control against overfitting
- ▶ No global optimum guarantee
(it could get stuck at the local maxima, saddle points, etc)
- ▶ We'll not determine number of components
➔ The initial values are important and several sets should be used