BROWN
Computer Science

CS1951A: Data Science

Lecture 8: Introduction to Hypothesis Testing

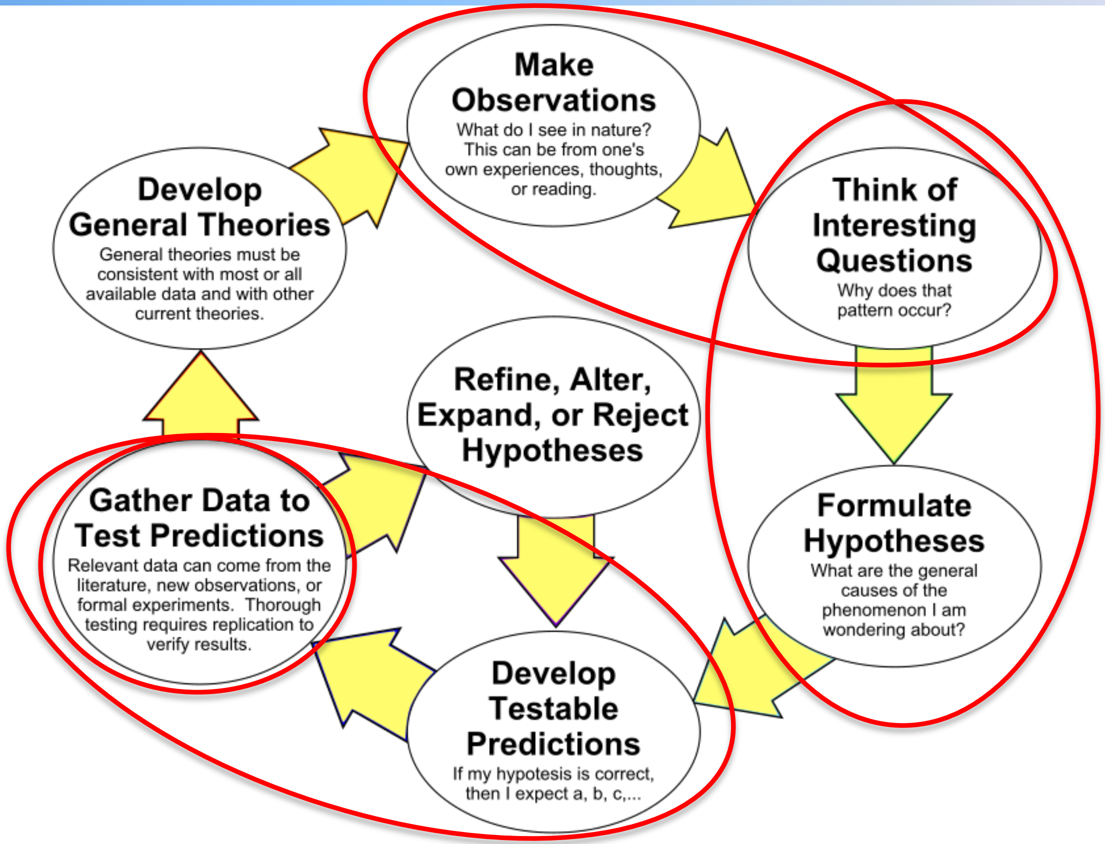
Lorenzo De Stefani

Spring 2022

Outline

- Approaches to data analysis: heuristic vs probabilistic vs statistical hypothesis testing
- What is a hypothesis
- Probability review
- A simple example of statistical testing

The Data Analysis Method



- “explore”, “analyze trends”, “look for patterns”, “visualize”
- Come up with possible explanations of the observed phenomena
  - Formulate hypotheses on the “world” from which the data is observed
- Test your hypotheses using new data from the same source
- Never use the same data to formulate hypotheses and to test them
  - Risk of overfitting and false discoveries

Approaches to data analysis: Heuristic

- Heuristic analysis:
  - We make observations on the available data
  - No (or very weak) guarantee on the generalizability of the results
  - Still can be very useful!
    - Some techniques within Machine Learning and Database analysis, and BigData analytics are heuristic in nature

Approaches to data analysis: Probabilistic method 1/2

- Probabilistic method:
  - We assume the existence of an underlying stochastic phenomenon
  - The phenomenon generates the observed data according to some unknown probability distribution (i.e., the ground truth)
  - We assume the observed data to be obtained by sampling such distribution (often assuming independently)
  - We analyze the data to infer conclusions that are valid for the entire underlying model
    - I.e., that generalize to the unknown distribution

Approaches to data analysis: Probabilistic method 2/2

- Probabilistic method:
  - We want guarantees on the accuracy of our insights!
    - if we are estimating a parameter we want to be able to claim that our estimate close to the true value (e.g., within ϵ<5%) with high probability (e.g., at least 99%)
  - Statistical machine learning, probabilistic data analysis, etc.,

Approaches to data analysis: Hypothesis testing 1/3

- We formulate a priori "null" hypothesis (H_0) on the "world" or some phenomena
  - You can think of this as some widely held belief
- We then come up with a new "alternative hypothesis" (H_1 or H_a) which contradicts the null.
  - This is the hypothesis we are interested in testing (our new belief)
  - It is not always complementary to the null
- We obtain some data to test our hypothesis

Approaches to data analysis: Hypothesis testing 2/3

- We do not directly test the alternative H_a
- Rather we test the null H_0
  - We will consider how unlikely it is that a phenomenon that follows the null hypothesis has generated data which behaves as the observed one or more extreme
  - Imagine this as saying:
    - Assuming that the null is correct and the phenomenon has given properties
    - How (un)likely am I to observe the given data!
- If the data appears extremely unlikely under the null hypothesis we reject the null hypothesis
  - We are saying the null hypothesis does not seem to correctly describe the phenomenon
  - CAREFUL: we are NOT accepting the alternative H_a
    - Rather we are saying H_a has a possibility of being correct
- Otherwise, we fail to reject the null hypothesis
  - CAREFUL: we are NOT accepting H_0

Approaches to data analysis: Hypothesis testing 3/3

- We want guarantees on the accuracy of our decisions:
- We would like to say that if we reject/fail to reject a null we are correct in doing so with some probability (confidence)
  - We will get asymptotic guarantees on the accuracy of our rejections
  - Very different from the probabilistic approach where we obtain finite sample guarantees
- Many important sub branches: classical "frequentist" statistical tests, Bayesian approach.

Statistics vs Probabilistic Analysis

- Both are useful and important
- There is a large intersection between the two
- Historical differences:
  - Back in the day smaller data were available
    - Focus on statistical hypothesis testing
  - In the era of BigData other methods are viable
- Depending on the available data one method may be more desirable
  - How much data is available?
  - How much prior information do I have about the model?
  - What kind of guarantees on the results do I want?
    - Statically testing yields asymptotic guarantees
    - Probabilistic analysis yields (stronger) finite sample guarantees

What is an hypothesis?

- A hypothesis is a statement about properties of an observed phenomenon
- It should be falsifiable
- It should be somewhat contested
  - Otherwise not very interesting!
  - Avoid tautologies

Quiz time!

“Look for differences in political affiliations between universities”

Is this a valid hypothesis?

a) Yes
b) No

Quiz time!

“Wearing a mask reduces the risk of contracting COVID”

Is this a valid hypothesis?

a) Yes
b) No

Quiz time!

“Wearing a mask reduces the risk of contracting COVID with respect to using no PID”

Is this a valid hypothesis?

a) Yes
b) No

What about these?

- h1: “This coin is biased towards head”
- h2: “People born in Europe are less likely to have chronic health conditions compared to people in East Asia”
- h3: “People with a college degree are more likely to enter the 1% of earners”
- h4: “Graduate students are disproportionally likely to being depressed”

The hypothesis testing method

1. Start from some observation on the data
2. Formulate a "research" (alternative) hypothesis H_a according to the prescribed rules
3. Test it against a "default" null-hypothesis H_0
4. Obtain fresh data to test the hypothesis
5. Using an opportunely chosen statistical test, determine if the data supports the null hypothesis or not

Probability spaces $< \Omega, F, P >$

A Probability Space has three components:

- A Sample Space Ω, which is the set of all possible outcomes of the random process being observed
  - E.g., Consider tossing a die we would have $\Omega = \{1,2,3,4,5,6\}$
  - E.g., Consider tossing two dice: what is Ω?

- A family of sets F representing the the allowable events, where each set in F is a subset of Ω
  - Elements of F also referred as "Events"
    - $F = 2^{\Omega}$
  - Elements of Ω referred as "Elementary events" or "Samples"
  - E.g., in our die example

$$F = \{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{1,2\}, \ldots \{1,2,3,4,5,6,\}\}$$

Probability spaces $< \Omega, F, Pr >$

- A probability function $\Pr: F \rightarrow [0,1]$ which satisfies the properties:
  - For any $E \in F$, $O \leq \Pr(E) \leq 1$
  - $\Pr(\Omega) = 1$
  - For any finite or countably infinite sequence of pairwise disjoint events $E_1, E_2, E_3, \ldots$

    $$\Pr(\cup E_i) = \Sigma \Pr(E_i)$$

    - E.g., for our die example

    $$\Pr(\{1\}) = \Pr(\{2\}) = \Pr(\{3\}) = \Pr(\{4\}) = \Pr(\{5\}) = \Pr(\{6\}) = 1/6$$
    $$\Pr(\{1,2,3,4,5,6\}) = 1$$

  - For any two events $E_1, E_2 \in \mathrm{F}$, $\Pr(\mathrm{E}_1 \cup E_2) = \Pr(\mathrm{E}_1) + \Pr(\mathrm{E}_2) - \Pr(\mathrm{E}_1 \cap E_2)$

Random Variables

- Sample space Ω: set of values which represent outcomes of an experiment
- A random variable X on a sample space Ω is a real-valued function on Ω, X: Ω→*R*
- A discrete random variable, is a random variable that can only assume a finite (countable) number of values.
  - The set of values it can assume is called the Range of the random variable
- Given a discrete random variable X and a real value a: the event "$X = a$" represents the subset of Ω given by $\{s \in \Omega : X(s) = a\}$

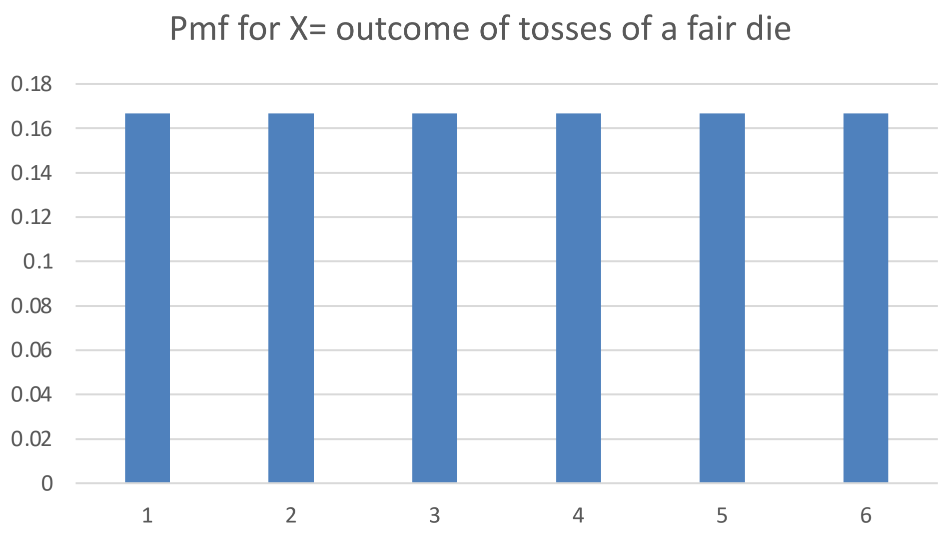
$$\Pr(X = a) = \Sigma_{s \in \Omega : X(s) = a} \Pr(s)$$

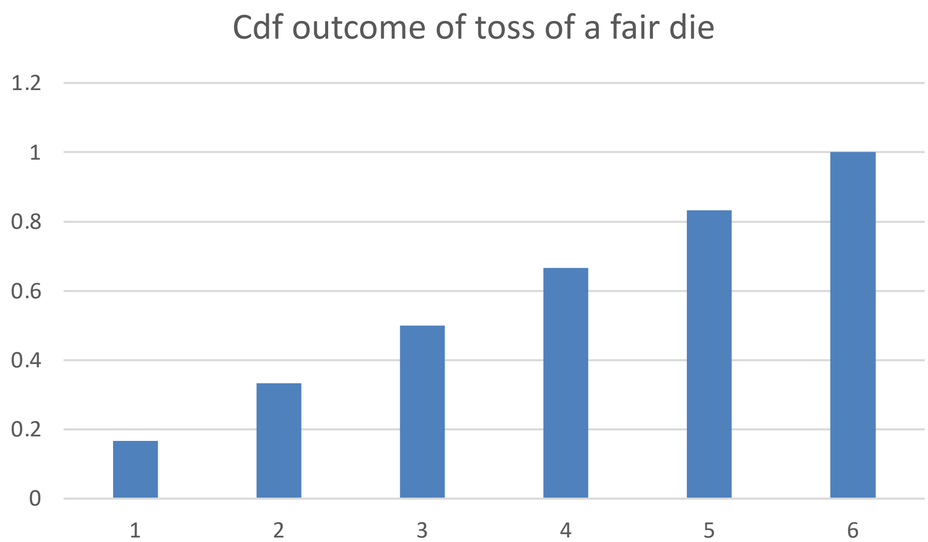
This is called the probability mass function of X (pmf)

- The cumulative distribution function (cdf) gives the probability of the random variable X assuming values up to a

$$\Pr(X \le a) = \Sigma_{a_i \le a} \Pr(X = a_i)$$

Example: fair die





Quiz time!

Let X be a random variable with the below cdf

X	1	2	3	4
$\Pr(X \leq a)$	0.5	0.75	0.9	1

What is the value of Pr(X≤3)?

a. 0.5
b. 0.15
c. 0.9
d. 1
e. 0

Quiz time!

Let X be a random variable with the below cdf

X	1	2	3	4
$\Pr(X \leq a)$	0.5	0.75	0.9	1

What is the value of Pr(X=3)?

a. 0.5
b. 0.15
c. 0.9
d. 1
e. 0

Quiz time!

Let X be a random variable with the below cdf

X	1	2	3	4
$\Pr(X \leq a)$	0.5	0.75	0.9	1

What is the value of Pr(X=2.5)?

a. 0.5
b. 0.15
c. 0.9
d. 1
e. 0

Independence

Two random variables X and Y are independent if and only if

$$\Pr\big((X = x) \cap (Y = y)\big) = \Pr(X = x)\Pr(Y = y)$$

for all values x and y. The definition extends to multiple random variables.

Identically Distributed RVs

Two random variables X and Y are identically distributed if and only if for all values x in the range of X and Y:

$$\Pr(X = x) = \Pr(Y = x)$$

- Sometimes we will say that two or more variables $X_1, X_2, \ldots X_i$ are iid as a short way to say that the variables are all
  - identically distributed and
  - pairwise independent

Expectation

- The expectation of a discrete random variable X, denoted as $E[X]$, is defined as

$$E[X] = \Sigma i \, Pr(\, X = i)$$

  where the summation is over the values i in the range of X.
  - $E[X]$ is a weighted sum over all possible values weighted according to their probability
- A continues random variable, is a random variable that can only assume an uncountable number of values (e.g., R)
- The expectation of a continuous random variable X, denoted as $E[X]$ is given by

$$E[X] = \int i \, P(X = i)$$

  where the integral is over the values i in the range of X.

Expectation

- The expectation is finite if it converges to a finite value, otherwise it is unbounded
- For any pair of random variables X_1, X_2 and constants a, b we have, by linearity of expectation

$$E[aX_1 + bX_2] = aE[X_1] + bE[X_2]$$

Expected Value

$$E(X) = \sum_i x_i Pr(x_i)$$

X	1	2	3	4
pdf	0.5	0.25	0.15	0.1

$$E[X] = ?$$

Expectation: example

Would you buy a bitcoin at \$30k knowing that it has a 10% chance to go to \$250k in a year, and 90% chance of going to \$0 in one year?

- Let us compute the expected value!
- X = gain with respect to the initial investment
  - pmf: Pr(X=220) = 0.1, Pr(X=-30)=0.9
  - E[X] = - 5k
- We expect to be losing money ☹

Expectation: example

- What should be different so that this in investment is worth considering?
- Expectation should be at least 0 (i.e., even odds), the higher the better
  - Higher chance of success/lower chance of loss
  - Higher return on success/lower loss of failure

Variance

- The variance of a random variable X, denoted as $\text{Var}[X]$ is defined as

$$\text{Var}[X] = E[X^2] - E[X]^2$$

- The standard deviation of X, is given by

$$\sigma[X] = \sqrt{Var}[X]$$

- They characterize how much deviation from the expectation we are likely to observe
- Very important for hypothesis testing!

Variance: Example

$$\text{Var}[X] = E[X^2] - E[X]^2$$

X	1	2	3	4
pmf	0.5	0.25	0.15	0.1

- Recall: $E[X] = 1.85$
  - $E[X]^2 = 3.4225$
- $E[X^2] = 1{\times}0.5 + 4{\times}0.25 + 9{\times}0.15 + 16{\times}0.1 = 4.45$
- $Var[x] = 4.45 - 3.4225 = 1.0275$

Variance: example

Would you buy a bitcoin at $30k knowing that it has a 10% chance to go to $250k in a year, and 90% chance of going to $0 in an year?

- E[X] = - 5k
- What is the variance?
  - $Var[X] = 495\times10^8 - 25\times10^6$
- Standard deviation?
  - $\sigma[X] \approx 225k$
- Both values are very high with respect to the expectation!
  - We are likely to observe large deviations from the expected value!
  - This is general undesirable when considering investments
  - Higher variance ➔Less predictability ☺

Binary random variables

A binary random variable can only assume two values

- If those values are 0 and 1 then it is called a Bernoulli random variable
- Used to represent many common phenomena
  - Coin tosses
  - Success/failure
  - On/off
  - ....
- The expectation of a binary RV is given by

$$E[X] = x_1 \Pr(X = x_1) + x_2 \Pr(X = x_2)$$

- Sometimes we say that a binary RV is fair or unbiased if the two outcomes have the same probability

Binomial distribution

Suppose we are flipping n times a coin and we want to characterize the probability distribution of the number of heads

- All coin tosses are independent of each other and identically distributed
  - We can model each coin toss as a Bernoulli RV X_i with probability of head = p
- We do not care about the order of heads and tails but only for the total final count!
- Let Y be the RV which denotes the number of heads

$$Y = \sum_{i=1}^{n} X_i$$

Binomial distribution

- Y is called a Binomial random variable with parameters (n, p)
  - n: number of attempts
  - p: probability of success in each attempt

- The pmf of Y is

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

$$P(Y = i) = \frac{n!}{i!\,(n - 1)!} p^i (1 - p)^{n-i}$$

  - The Binomial coefficient $\binom{n}{i}$ counts all the possible outcome sequences with i successes and $n - i$ failures

Expectation of known distributions

- Let X be a Bernoulli random variable with parameter p

$$E[X] = ?$$

- Let Y be a Binomial random variable with parameters n, p

$$E[Y] = ?$$

Variance of known distributions

- Let X be a Bernoulli random variable with parameter p

$$\text{Var}[X] = ?$$

- Let Y be a Binomial random variable with parameters n, p

$$\text{Var}[Y] = ?$$

The bigger picture

- Start with real world/phenomenon observations
- Make assumptions about the underlying model
  - Se the null hypothesis
- Fit the parameters of the model based on data
  - Chose parameters of the model based on theories, do analysis to see if its a good fit (hypothesis testing!!)
  - Set parameters of the model based on data, try to make forecast for unseen/future data (prediction!!)

Hidden patterns in driving license tests

Are the correct answers to multiple choices quiz truly random?

- Null hypothesis h: "I think that for each question the answer "b" with 80% probability"
- To test our hypothesis we collect some data:

a	b	c	b
a	b	c	c
d	c	b	d

- What is the likelihood of observing such data assuming that the hypothesis is correct?

Hidden patterns in driving license tests

What is the likelihood of observing such data assuming that the hypothesis is correct?

a	b	c	b
a	b	c	c
d	c	b	d

- We define the probability space for each question under the current hypothesis
  - $\Omega = \{b,\ not\ b\}$
  - Pr(b) =0.8, Pr(not b) = 0.2
- We are also implicitly assuming that the questions are independent and identically distributed

The probability of observing such data under the current assumption is = $0.8^4 \times 0.2^8 = 0.00000105$

- This seems very low....so we can for sure say that the hypothesis is not correct....right???
- NOT QUITE SO FAST ☺

Hidden patterns in driving license tests

What if we consider a different set of data?

a	b	b	b
b	b	b	c
b	b	b	b

- We define the probability space for each question under the current hypotheis
  - $\Omega = \{b,\ not\ B\}$
  - Pr(b) =0.8, Pr(not B) = 0.2
- We are also implicitly assuming that the questions are independent and uniformly distributed

The probability of observing such data under the current assumption is then = $0.8^{10} \times 0.2^{2} = 0.004$

- But this still seems low even though the data seems to strongly support the hypothesis
- Are we doing something wrong???

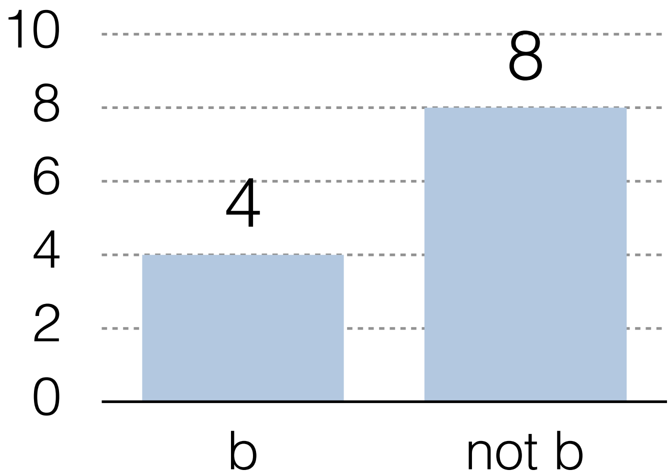
A different question

The error is in the question that we ask and the ways we interpret its result

- The absolute probability of an event is not by itself decisive
- Rather than just asking how likely it is to observe the data, we should ask how likely it is to observe something that looks much different than this!

Revisiting the question

Rather than considering a specific order of answers, we focus on the aggregate distribution of the answers



The random phenomenon we care about is the number of questions with answer b

$$X = \ questions\ with\ answer\ B$$

- X is a random variable
- What is its distribution?

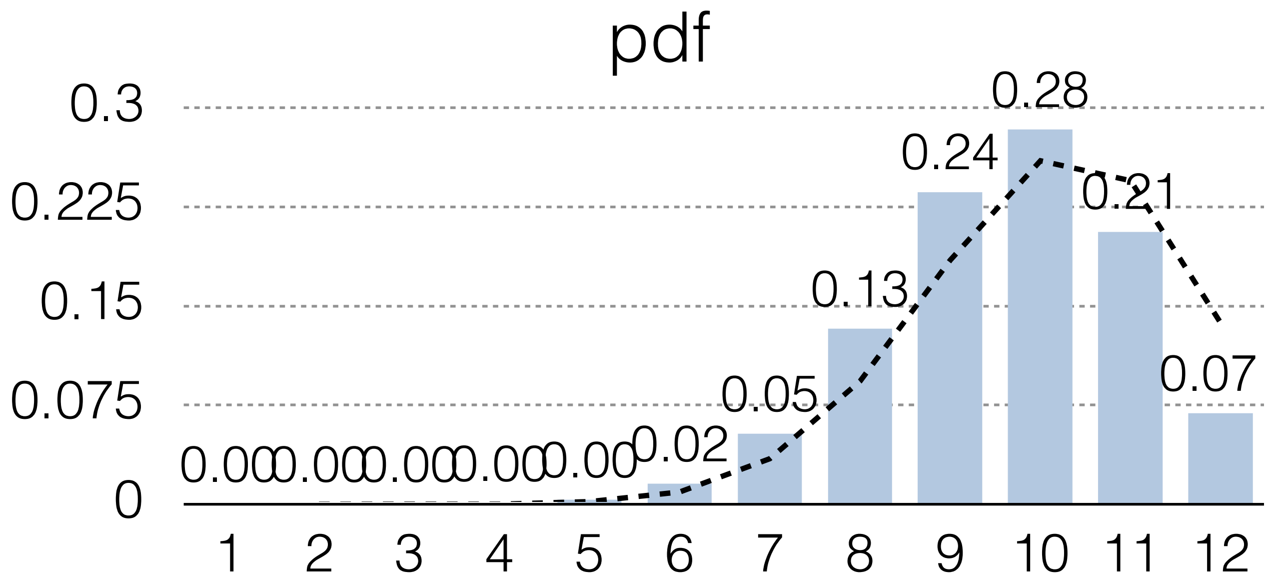
According to our hypothesis, X is the sum of random variables, each corresponding to each question, whose answer is b with probability 0.8

- X is Binomial random variable with parameters 12,0.8

Binomial distribution of the outcomes

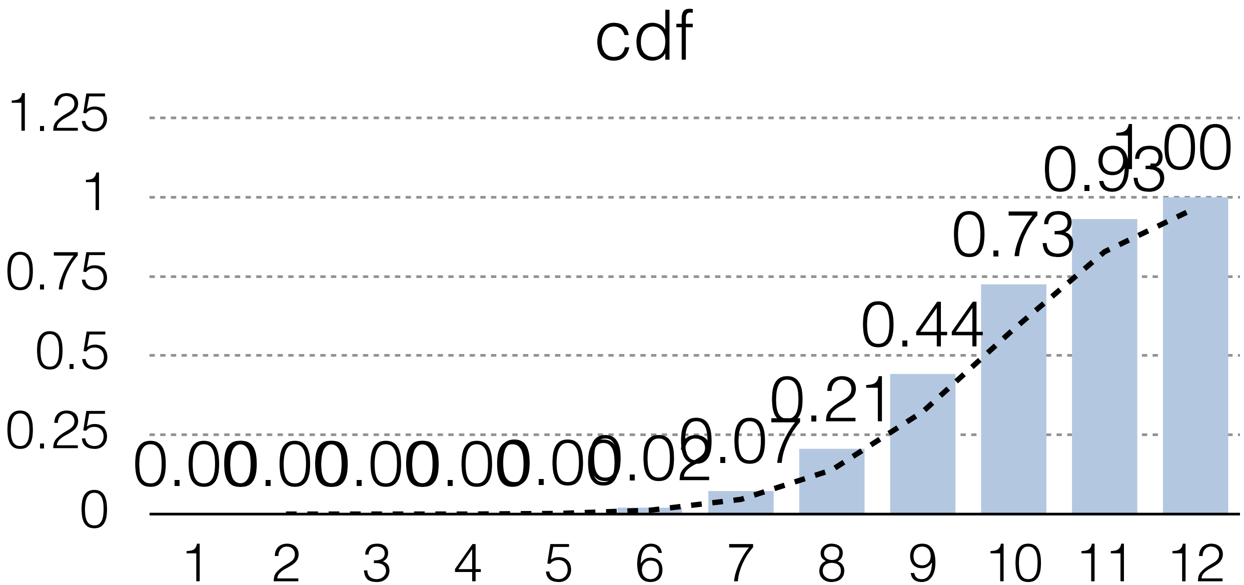
h: “I think that for each question the answer “b” with 80% probability”

$$P(Y = i) = \binom{12}{i} 0.8^{i} 0.2^{12-i}$$



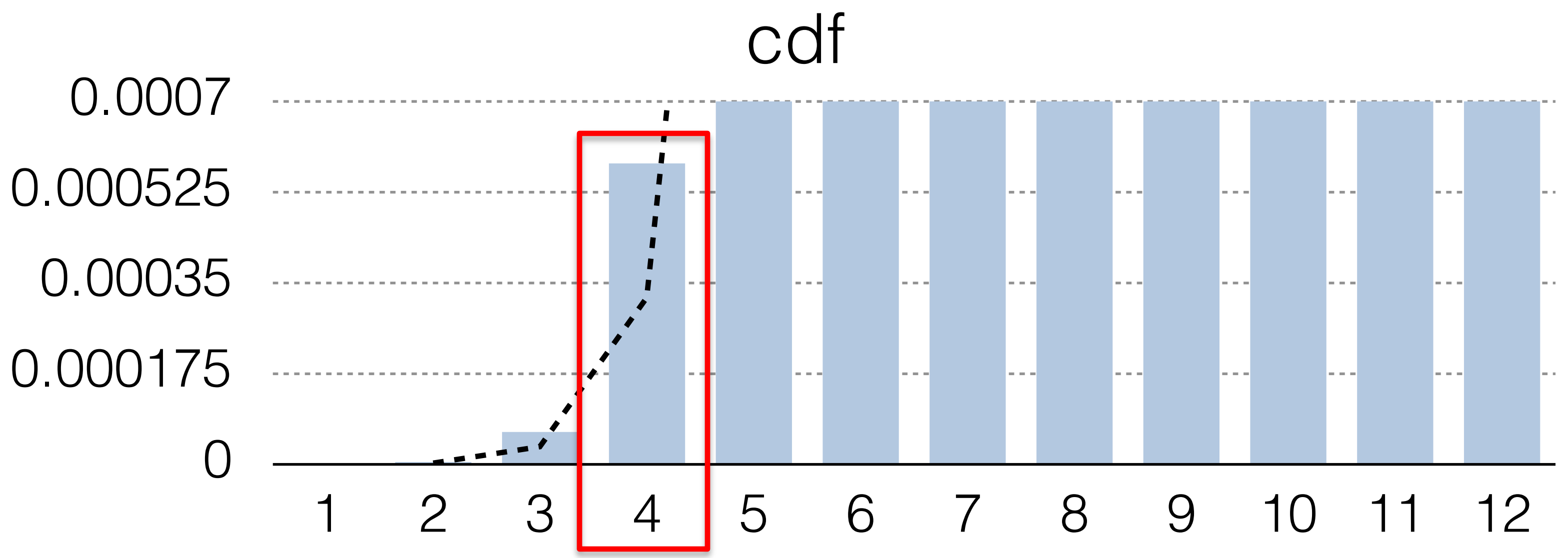
Binomial distribution of the outcomes

h: "I think that for each question the answer "b" with 80% probability"



Is the fact that we observed 4 b very unlikely under the current assumption on the model?

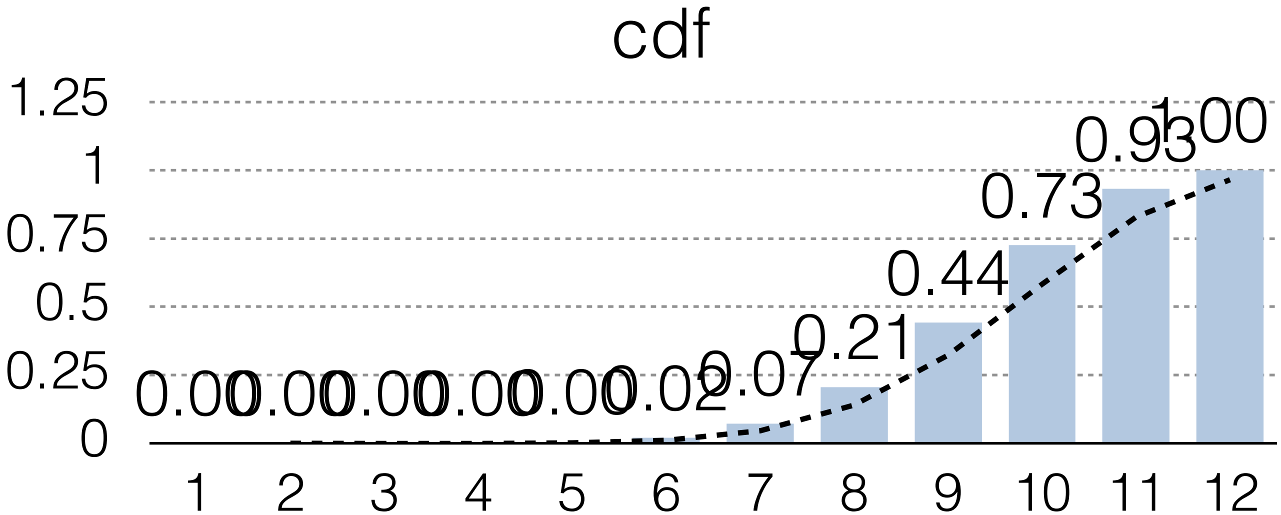
Does the data support my hypothesis?



$P(X \leq 4) < 0.0006!$

- The probability of the observed data under the current model is very low
- We would be incline to reject the hypothesis as unlikely to be correct!

Does the data support my hypothesis?



$P(X \leq 10) < 0.73!$

- The probability of observing no more than 10 b's is rather high
- The data appears to support the hypothesis
  - We do not have evidence to reject it
  - CAREFUL: we are not saying that the null hypothesis is very likely correct

To be continued

- Over the next few lectures, we will introduce statistical tests which give us a way to measure how much the data "supports our hypothesis"
- We will introduce the idea of p-value which gives us a criteria for deciding which hypotheses can be rejected with some guarantee

Conclusion

h: “I think that for each question the answer “b” with 80% probability”

At the end of the day, was this hypothesis correct?

a) yes
b) no