Questions from a student

Tutorium 1:

-> **Folie 29:** "Die Wahrscheinlichkeit eines Ereignisses ist die Summe der Wahrscheinlichkeiten der entsprechenden Ergebnisse" Könntest Du mich mal, bitte, Anhang dieser Folie noch mal den Unterschier zwischen einem Ereignis und einer Ergebnis erklären.

-> **Folie 41:**

"compute • $P(X=0)$ $P(X=0) = p(schlecht)$" Was es bedeutet und worum geht es hier?

Tutorium 2:

-> **Folie2:** Was ist einzelne Entropie und wie berechnet man sie?

There are mistakes in page 29 and 41, please download the new version of the slide

Materialien

Date	Slide	Note
29.10.2021	tutorium_1_corrected.pdf	corrected mistakes on page 29, 40, 41
5.11.2021	slide2.pdf	Erwartungswert, Varianz, Binomialtest
......	......	......

These answers are also in the Fragensammlung

Tutorium 1:

-> **Folie 29:** "Die Wahrscheinlichkeit eines Ereignisses ist die Summe der Wahrscheinlichkeiten der entsprechenden Ergebnisse" Könntest Du mich mal, bitte, Anhang dieser Folie noch mal den Unterschier zwischen einem Ereignis und einer Ergebnis erklären.

Zufallsexperiment: Experiment (Versuch) mit mehreren möglichen Ausgängen (*Wurf von zwei Würfeln*)

Ergebnis: Resultat eines Experimentes (*3 Augen auf Würfel 1 und 4 Augen auf Würfel 2*)

Ergebnisraum Ω: Menge aller möglichen Ergebnisse

Ereignis $A \subseteq \Omega$: Teilmenge des Ergebnisraumes (*7 Augen mit 2 Würfeln*)

s.16

Ein Ergebnis(outcome) ist ein Resultat des Zufallsexperimentes.
Beispiel: Wir werfen einen Würfel und betrachten die Augenzahl des Würfels. In diesem Fall gibt es 6 Ergebnisse, nämlich die Augenzahl 1, 2, 3, 4, 5, 6.

Augenzahl1 ist ein Ergebnis.
Augenzahl2 ist ein Ergebnis. Und so weiter.

Wenn man alle möglichen Ergebnisse in eine Menge packt, dann hat man den Ergebnisraum.
Ergebnisraum = {Augen1, Augen2, Augen3, Augen4, Augen5, Augen6 } oder wir schreiben das einfach als {1,2,3,4,5,6}

Ein Ereignis(Event) ist eine Teilmenge des Ergebnisraumes.
Z.b. {1,2,3} oder {1} oder {2,4,6}, oder das ganze {1,2,3,4,5,6} sind alle ein Ereignis.

Folie 29

Die Wahrscheinlichkeit eines Ereignisses ist die Summe der
Wahrscheinlichkeiten der entsprechenden Ergebnisse.

Fehler: hier soll "Ereignis" stehen

- Was brauchen wir, um die WK eines Ergebnisses zu berechen?
  - Wir müssen wissen, welche Ergebnisse in diesem Ergebnis enthalten sind. Z.b. A = {2, 4, 6 }
  - Wir müssen die WK für jedes Ergebnis kennen, also p(2), p(4), p(6)

- Wir können dann P(A) berechnen
  - P(A) = p(2) + p(4) + p(6)

o steht für ein Ergebnis aus A

$$\sum_{o \in A} p(o)$$

"compute • P(X= 0) P(X=0) = p(schlecht)" Was es bedeutet und worum geht es hier?

0.25
gut
ok   0.25
schlecht
0.5

1

0
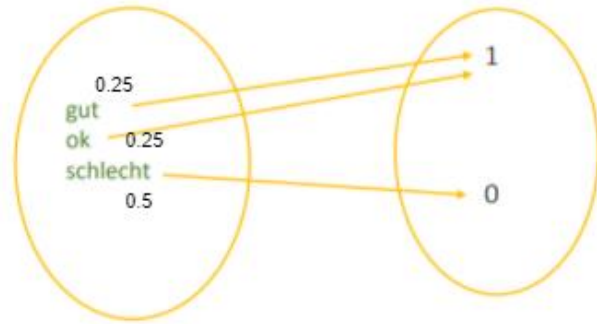
P({gut, ok, schlect}) = 1

Steps
- what is X=1 ?
  - it is the set { gut, ok, ~~schlecht~~}
  - this is an event and to compute the probability of an event we have to compute
    - the sum of each outcome in that event
    - P(X=1) = p(gut) + p(ok) + ~~p(schlecht)~~

Hier wird gezeigt, wie man P(X=1) berechnet

- compute P(X= 0)
  - P(X=0) = p(schlecht)

Hier möchte ich P(X=0) berechnen.
X=0 entspricht der Menge {schlecht}, deswegen wird p({schlecht}) berechnet. Es ist auch OK, ohne { } zu schreiben.

Gut, OK, Schlecht sind die Ergebnisse von einem Experiment. In diesem Beispiel ist das Experiment eine Beobachtung von irgendeinen Bewertungen/Noten. Wir nehmen hier an, es gibt nur 3 Noten. Wir definieren dann eine Zufallsvariable X mit folgender Zuordnung.
Wir ordnen den Ergebnissen "gut" und "ok" die Zahl 1 zu.
Es wird als X=1 notiert. X=1 repräsentiert die Menge {gut, ok}.
X=0 repräsentiert die Menge {schlecht}

Tutorium 2:

-> **Folie2:** Was ist einzelne Entropie und wie berechnet man sie?

- Was ist einzelne Entropie

"Einzelne Entropie" ist H(Y) gemeint, was "Entropie der Zufallsvariable Y" genannt werden soll.

- Wie berechnet man sie

Beispiel
Gegeben eine Zufallsvariable Y mit
Y=1 mit P(Y=1) oder p(1) = 0.4
Y=2 mit P(Y=2) oder p(2) = 0.6

H(Y) = - [p(1) log_2 p(1) + p(2) log_2 p(2)]

Die **Entropie** misst, wieviel Information ein Zufallsereignis im Mittel enthält.

Entropie einer Zufallsvariablen X mit der Verteilungsfunktion $p(x)$:

$$H(X) = H(p) = -\sum_{x \in X} p(x) \log_2 p(x) = E\left(\log_2 \frac{1}{p(x)}\right)$$

Wie verwenden hier diese
Formula

Binomial distribution and Binomial test

Binomial distribution

- It is a probability distribution that can be defined when we have a Bernoulli experiment.
- Example: We have a coin toss experiment with 2 outcomes (head, tail) and we define a random variable X with X=1 for head (also called **success event**) and X=0 for tail. This is a Bernoulli experiment.
- Now we want to know how many times we are most likely to get head if we repeat the experiment 100 times.
- Or the question, "Is it more likely to get a head 50 times than 20?".

- If we already know the probability of getting a head P(X=1), then we can use a binomial distribution to answer these questions without having to run the actual experiment.
- For example, if we set n=100, p=0.5 (the probability of getting head), we can plot the graph of a binomial distribution as follows.
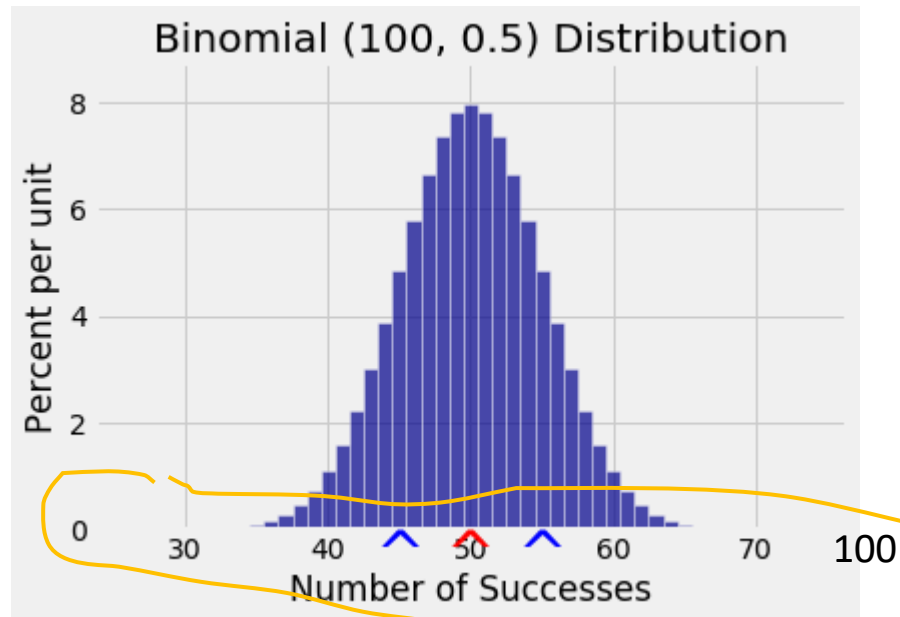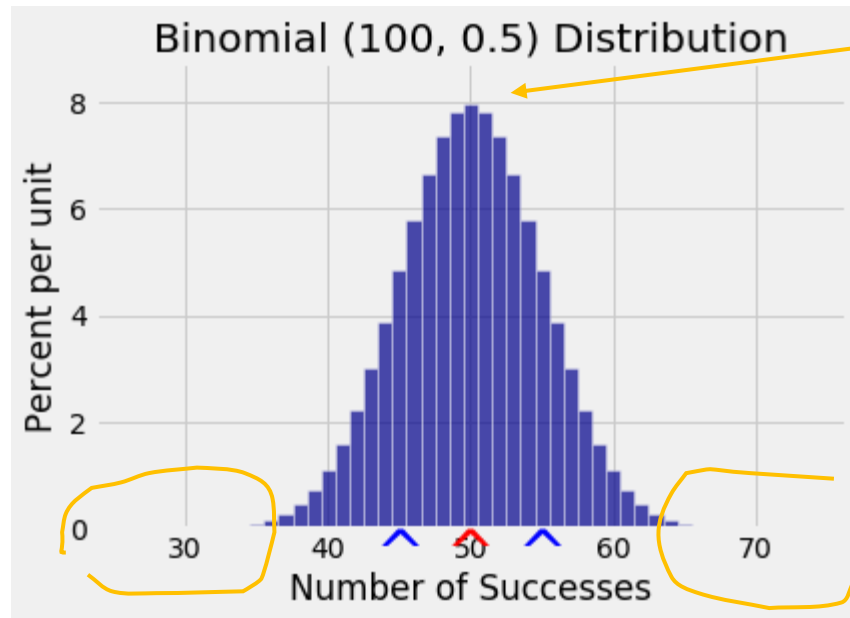

The value here will be 0...100.
This is the value of a new random variable for which we define the binomial distribution. Let's call it B.
P(B=0) is the probability that we get head zero time out of 100.
P(B=1) is the probability that we get head 1 time out of 100.
...
P(B=100) is the probability that we get head 100 time out of 100.

Binomial (100, 0.5) Distribution

We can see that 50 times has the highest probability. This means that it is most likely to get head 50 times.

And it is very unlikely to get 0 or 100 times head.

If P(head) is not 0.5, but higher, for example, p(head) = 0.7 The binomial graph might look different.
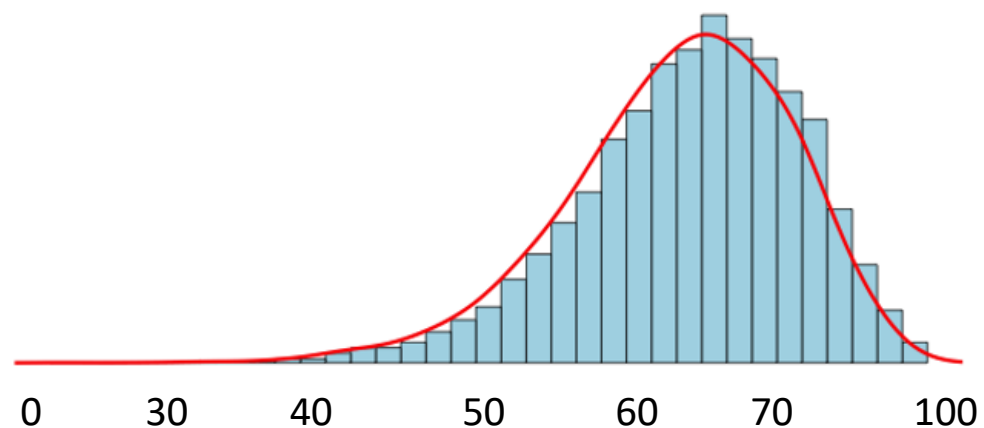
Bernoulli experiment.

- experiment: tossing a coin, observe if you get head or tail

- Sample space = {head, tail}
- X=1 ; {head}
- X=0; {tail }

Binomial

- Experiment: repeat the coin toss experiment 100 times and observe number of times we get head
- Sample space= {0,1,2,3..., 100}
- B=0 ; {0}
- B=1 ; {1}
- ...
- B=100 ; {100}

- * there is a formula to compute P_binomial if you know p, r, n
  - r = the number of time that head occurs which we are interested in
  - n = number of time we repeat the experiment
  - p = probability of head

$$b(r; n, p) = \binom{n}{r} p^r (1 - p)^{n-r}$$

$$\binom{n}{r} = \frac{n!}{(n - r)!\, r!}$$

Formula for calculating a binomial probability

If you compute b(r=50, n=100, p=0.5) you will get the probability mass of getting head 50 times

$$b(r; n, p) \ = \ \binom{n}{r} p^r (1-p)^{n-r}$$

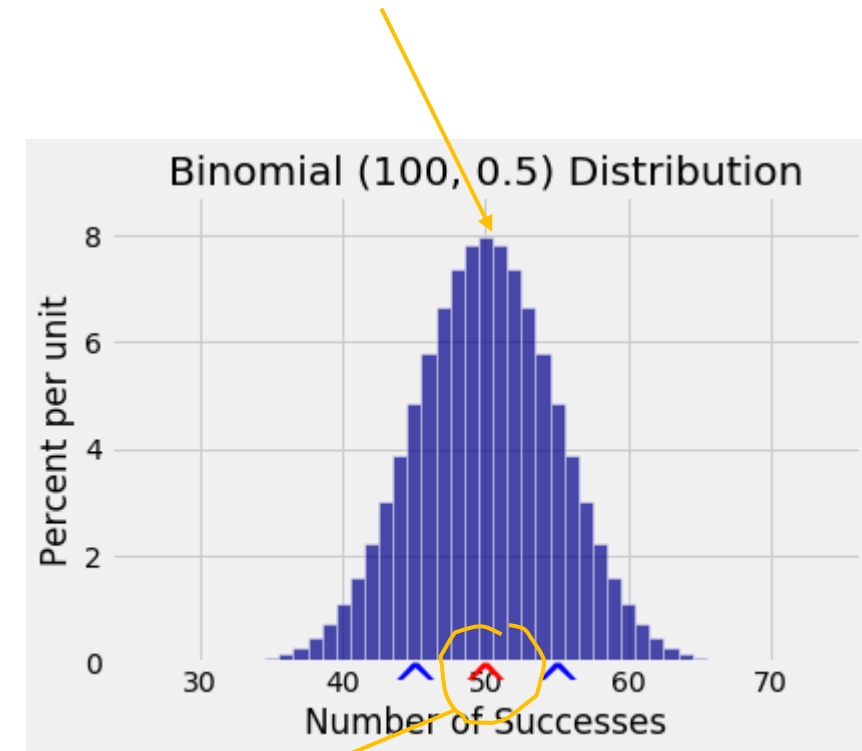$$\binom{n}{r} \ = \ \frac{n!}{(n-r)! \; r!}$$

We need r, n, p to compute this.
r is the number of time that head occurs
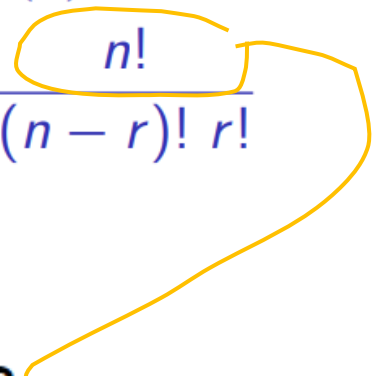n is the number of time we repeat the experiment
p is the probability of getting head

If we have n=100, r will be the number between 0 and 100.

Binomial (100, 0.5) Distribution

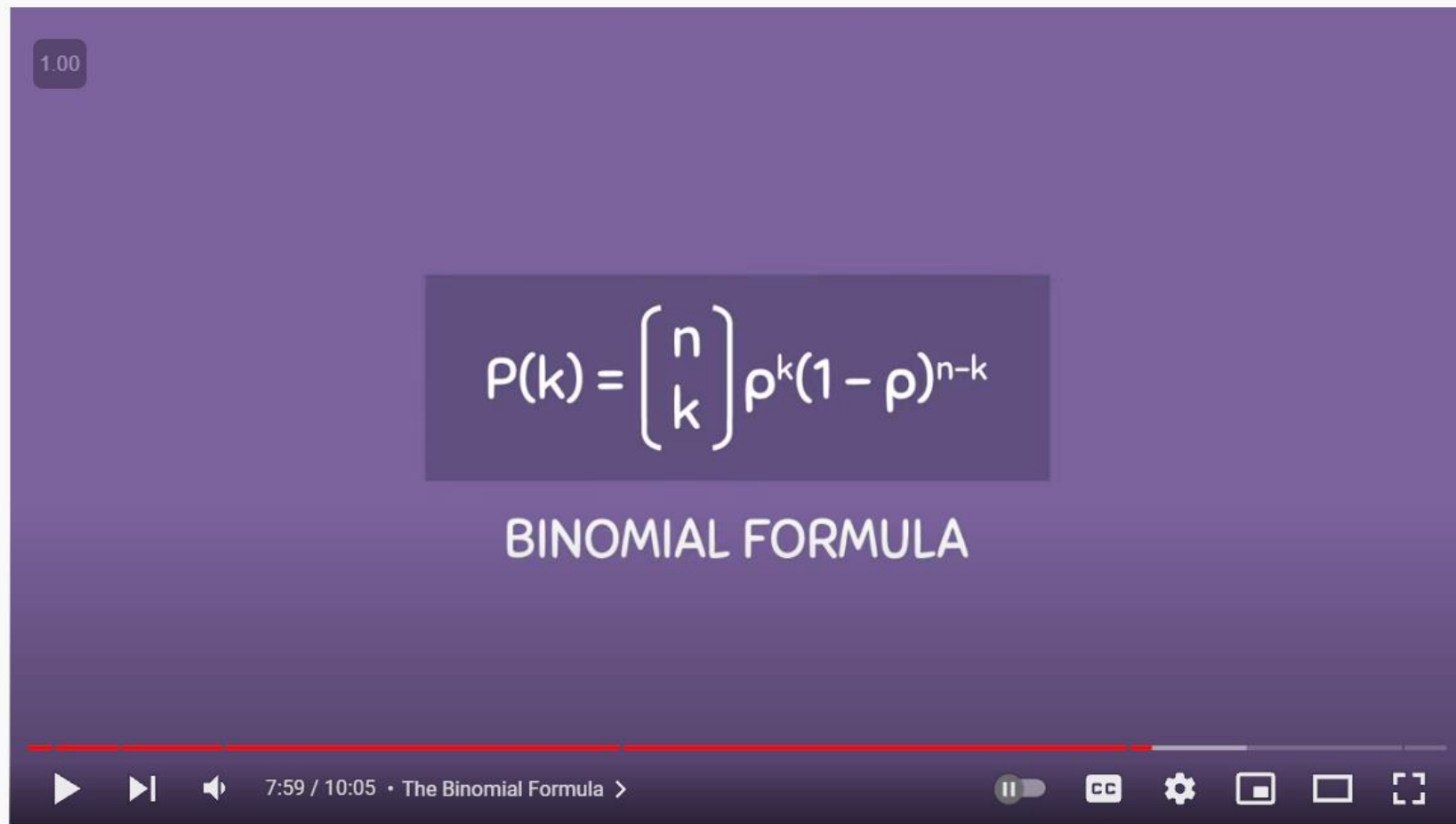Percent per unit

Number of Successes

r

$$b(r; n, p) = \binom{n}{r} p^r (1 - p)^{n-r}$$

$$\binom{n}{r} = \frac{n!}{(n - r)! \, r!}$$

What is a Factorial?

A factorial is a function that multiplies a number by every number below it. For example 5!= 5*4*3*2*1=120. The function is used, among other things, to find the number of way "n" objects can be arranged.

The Binomial Experiment and the Binomial Formula (6.5)

https://www.youtube.com/watch?v=nRuQAtajJYk&t=94s&ab_channel=SimpleLearningPro

$$P(k) = P(k\ successes) = \binom{n}{k} p^k\ q^{n-k}$$

$$P(k) = \frac{n!}{k!\,(n-k)!}\ p^k\ q^{n-k}$$

Binomial Distribution Probablity Coins

1:04 / 7:51

https://www.youtube.com/watch?v=c2gvvg_zxWQ&ab_channel=statisticsfun

Statistical hypothesis testing

Situation: You know that if the coin is fair, p(heads) must be 0.5, and if you plot a binomial distribution, you would get the highest probability for heads at 50 times. This is what you would expect.

Now you have done the actual experiment by flipping a coin 100 times and getting heads = 80, tails = 20 times. You find that the result is strange because the number of times you get heads is very far from what you expect.

Binomial (100, 0.5) Distribution
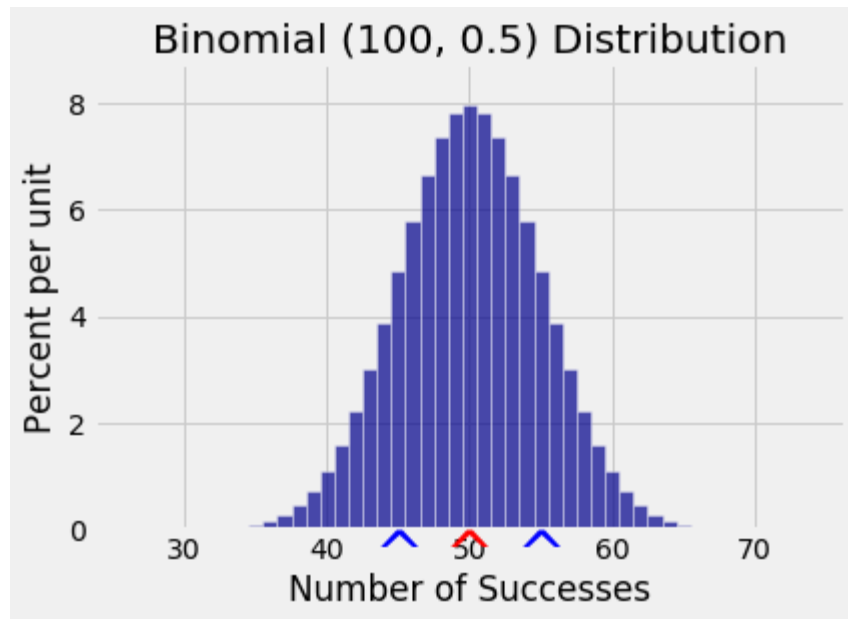
So you wonder if p=0.5 is still correct? Is your coin biased? Or is it just a coincidence that you get this strange result. Maybe this sample space has a different distribution and it's not P(heads)=0.5 and P(tails)=0.5. Maybe P(heads) > 0.5.

However, based on your experiment, you cannot draw a conclusion about p yet. This is because n= 100, which you selected, may not be large enough to make any statement about p.

And even though the probability of getting head 80 times is very low in this situation, it is not zero. This means that there is still a small chance to get such a result even if p(head)=0.5.
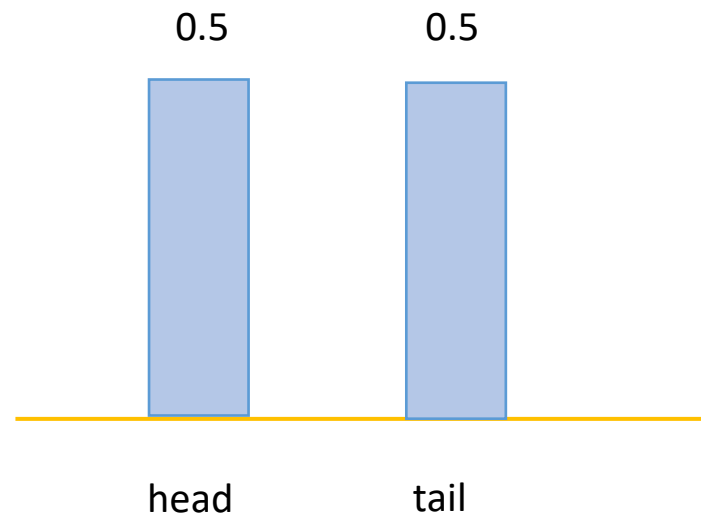
Hypothesis testing can be used to determine if the results (heads = 80, tails = 20, n=100) are significant enough to support your assumption (that the coin has a tendency to go heads).

In this case, we will **use the binomial test.**

the binomial test: Set the null hypothesis

H_0 is what you want to reject

This is what you believe/assume/suspect

H_0 : null hypothesis saying p(head) = 0.5

You suspect that p(head) might be greater than 0.5
H_alternative: p(head) > 0.5

0.5   0.5

> 0.5

< 0.5

head   tail

head   tail

the binomial test

100 Coin Tosses
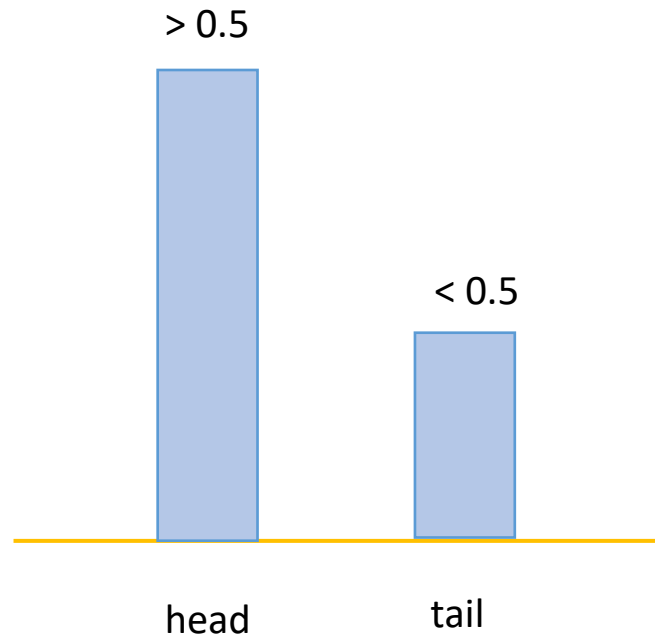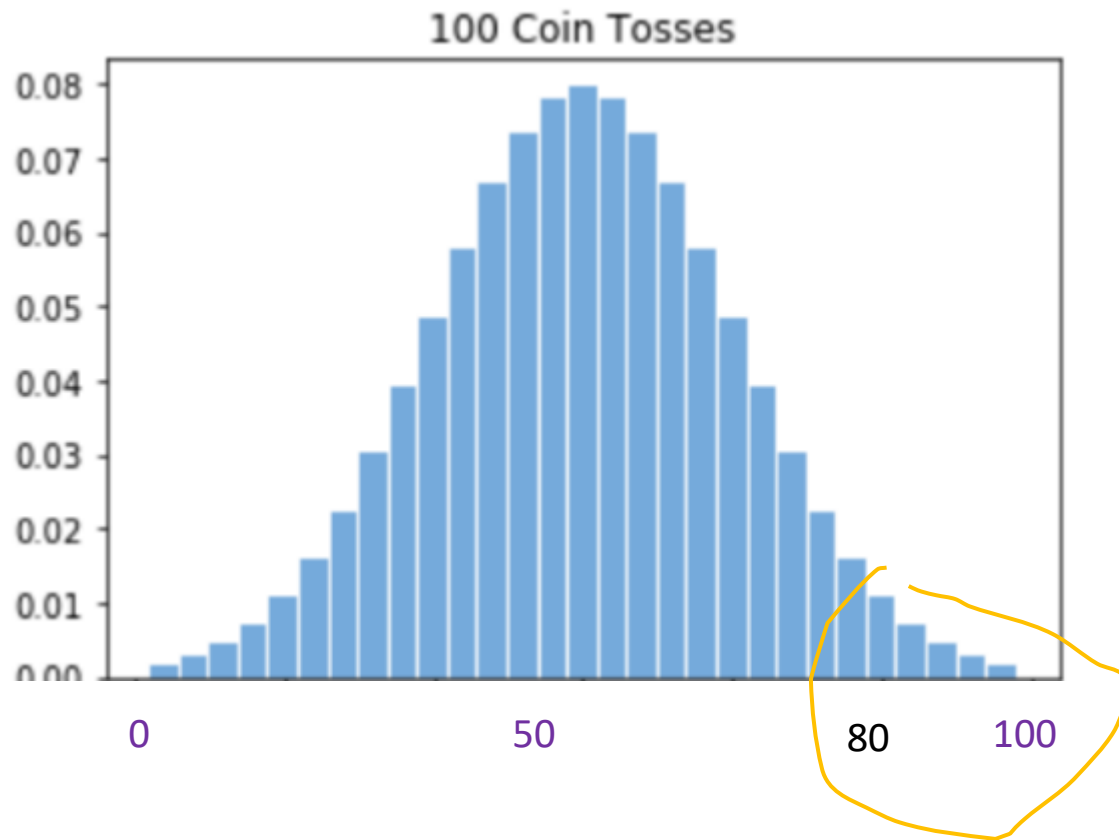
$$P(B>=80) = P(B=80) + P(B=81) + P(B=82) +...+ P(B=100)$$

The idea is to see if the probability of getting head 80 times to 100 times is low enough to say that the experiment might have a different distribution than what is defined in H_0. In other words, is it low enough to say that p(head) is not 0.5 but more? This probability is written as P(B>= 80) or P(>= 80).

You can think of it this way, if P(B>= 80) is super low, it means getting head 80 times does not seem like a normal thing to occur under the null hypothesis which says p(head)=0.5. This will give us a reason to reject the null hypothesis.

P(B>= 80) will not be zero, but if it is very near zero, we can assume that p might not be 0.5, but something else.

(How low is low enough?)
Now you have to pick a significant value. If you pick 0.05 it means you will only believe your alternative hypothesis (which says p(head)>0.5) only if P(B>= 80) is less than 0.05.
The significant value is commonly picked as 0.05 or 0.03, or 0.01.
(The lower, the more trustworthy is your test result)

For example, if P(B>= 80) = 0.0123 which is less than 0.05. You can reject the null hypothesis.
If p(B>= 80) = 0.0567, you can **not** reject the null hypothesis, because the probability is greater than 0.05.
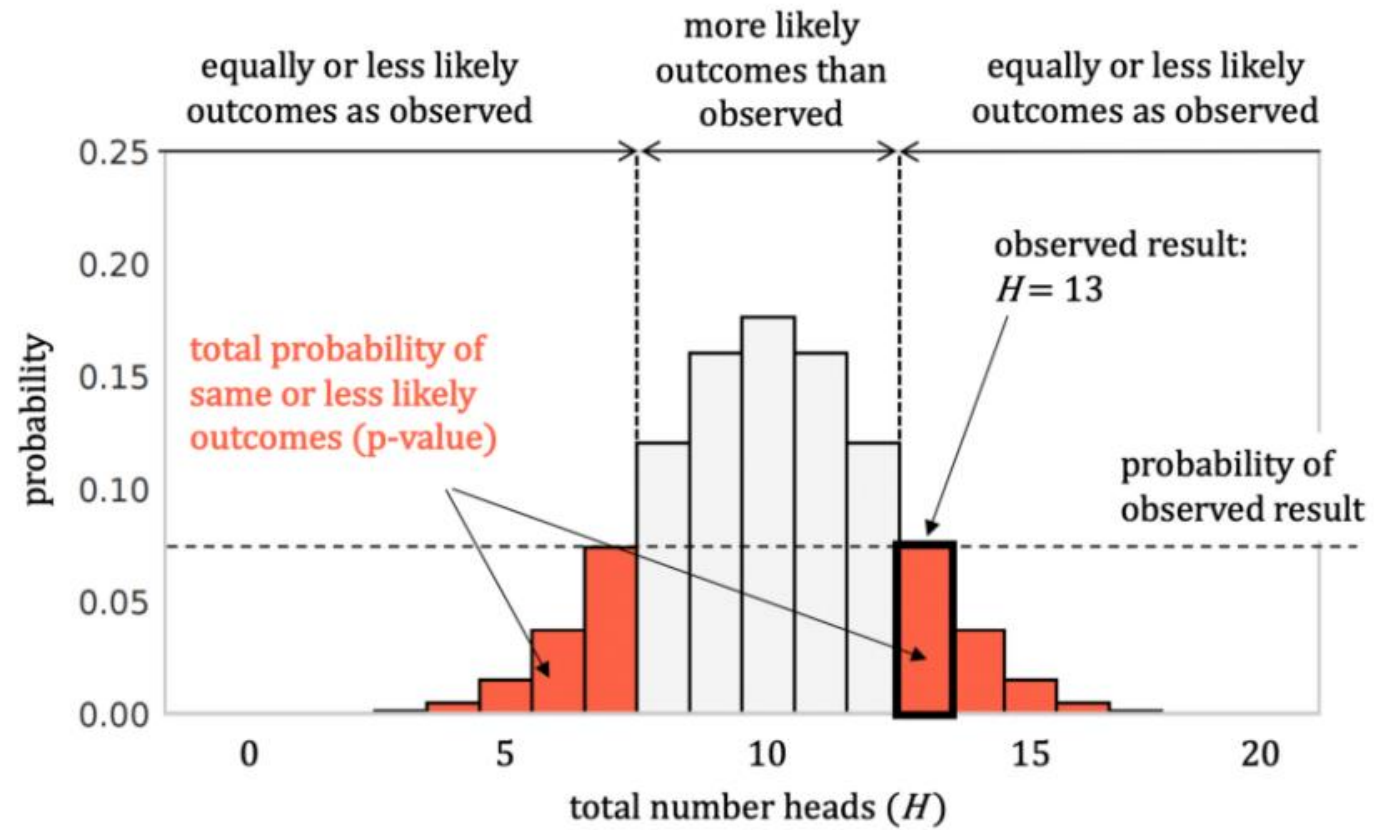
Figure 1. Probability distribution for all possible outcomes of 20 coin flips assuming the coin is fair and P(heads) = 0.5. Function is represented by Eq. 1 with parameters Bin(H, Ntosses = 20, p = 0.5). The area in red corresponds to the total probability of observing an equally or less likely outcome and is called p-value.

Exercise:

1. You toss a coin 80 times and get head 10 times and tail 70 times. Is the coin biased? and biased toward head or tail? Prove this using the binomial test.

2. You toss a coin 10 times and get head 8 times and tail 2 times. Is the coin biased? and biased toward head or tail? Prove this using the binomial test.

100 Coin Tosses
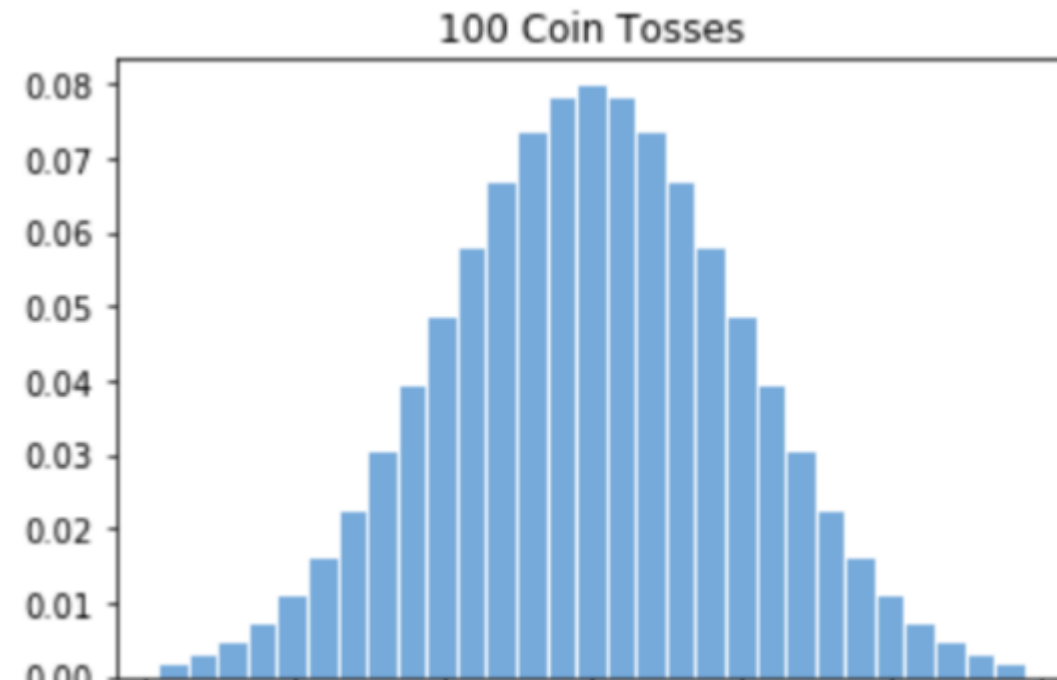
Solution:

1. You toss a coin 80 times and get head 10 times and tail 70 times. Is the coin biased? and bias toward head or tail? Prove this using the binomial test.

- We suspect that the coin is biased in favor to tail, so we set H_Alternative: P(head) < 0.5
- We then have to set the null hypothesis to be the opposite of what we believe, namely, the coin is not biased. Which means P(head) = 0.5

H_0 : P(head) = 0.5
H_Alternative: P(head) < 0.5

Pick a significant value = 0.05   (you can pick 0.03 or 0.01 as well)

Why do we compute P(X<=10) ? The idea is to use the actual observation to help find out if p defined in H_0 is correct (can not be rejected).
We get 10 heads from our actual experiment and if H_0 is correct (meaning p(head) = 0.5) we expect that P(X<=10) should be a quite high probability, indicating that it is a normal thing that occurs under this p.
And if it is not the case, e.g. P(X<=10) turns out to be super low, then we can reject H_0.
We compute not only X=10 but also X=9 and lower, because they support our assumption, p(head) < 0.5, even more.

compute P_binomial(X<= 10) which is P_binomial(X=0) + P_binomial(X=1)+ P_binomial(X=2) + ... + P_binomial(X=10)

$= b(r=0, n=80, p=0.5) + b(r=1, n=80, p=0.5) + b(r=2, n=80, p=0.5) +... b(r=10, n=80, p=0.5)$

100 Coin Tosses
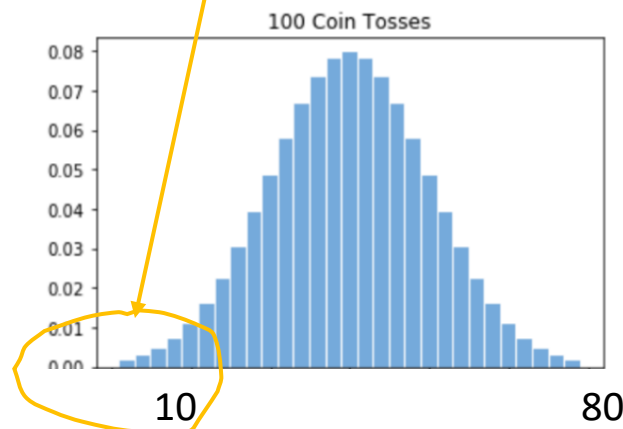
If P_binomial(X<= 10) < 0.05, then we reject H_0 (meaning the test suggests that the coin is biased toward tail), otherwise we accept H_0 (meaning the test suggests that the coin is not biased).

note: In the exam, you don't have to actually compute **P_binomial(X<= 10)** because it would take too long. It is enough to show how to compute it.

http://statisticshelper.com/binomial-probability-calculator

the result is a lot lower than 0.05, so we can reject H_0

Binomial Probability Calculator

Trials (n): *

80

Probability (p): *

0.5

Successes (X): *

10

Type of probability:*

At most X successes

CALCULATE

Answer:

$P(X \leq 10)$ Probability of at most 10 successes: 1.5806452613247E-12

Solution:

$P(X \leq 10)$ **Probability of at most 10 successes**

At most 10 successes includes X-values of $X = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$. To solve this problem, find the sum of the binomial probabilities for each of the values of X, or if there is only one value of X, find the probability of P(X). In this problem,

$$P(0) + P(1) + P(2) + P(3) + P(4) + P(5) + P(6) + P(7) + P(8) + P(9) + P(10)$$

Solution:

1. You toss a coin 10 times and get head 8 times and tail 2 times. Is the coin bias? and bias toward head or tail? Prove this using the binomial test.

- We suspect that the coin is biased in favor to head, so we set H_Alternative: P(head) > 0.5
- We then have to set the null hypothesis to be the opposite of what we believe, namely, the coin is not biased. Which means P(head) = 0.5

H_0 : P(head) = 0.5
H_Alternative: P(head) > 0.5
Pick a significant value = 0.05 (you can pick 0.03 or 0.01 as well)

compute P_binomial(X>=8)

      P_binomial(X>=8) = P_binomial(X=8) + P_binomial(X=9)+ P_binomial(X=10)

100 Coin Tosses
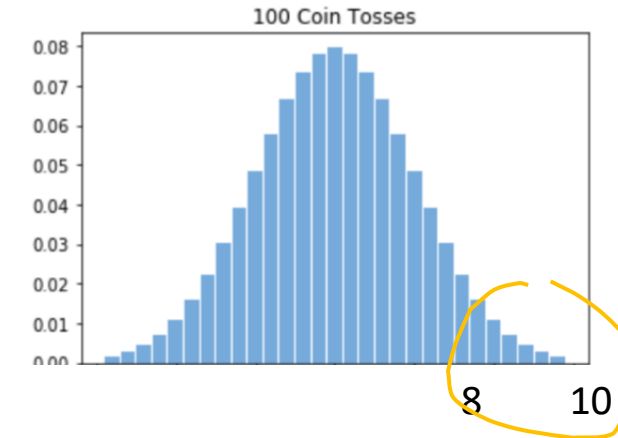
If P_binomial(>=8) < 0.05, then we reject H_null (meaning the test suggests that the coin is biased toward head), otherwise we accept H_null (meaning the test suggests that the coin is not biased).

Why do we compute P(X>=8) and not P(X<=8) or something else ?
8 is the actual observation of getting head. If H_0 is correct, meaning p(head) = 0.5, then the binomial distribution should give us a high value for P(X>=8), which indicates that it is a normal thing to occur under p(head) = 0.5.
We not only compute P(X=8) but also value >=8 because of our assumption. We believe p(head) > 0.5. If we get head more than 8, for example, 9 or 10. It will even be a strong indicator that P(head) > 0.5. This is why we compute P(X>=8) and not P(X<=8) or something else.

Binomial Probability Calculator

Trials (n): *

10

Probability (p): *

0.5

Successes (X): *

8

Type of probability:*

At least X successes

CALCULATE

it is greater than 0.05, so we can not reject H_0

Answer:

$P(X \geq 8)$ Probability of at least 8 successes: 0.0546875

Kollokationen

Kollokationen sind feste Wortkombinationen, die beim Erwerb einer Sprache gelernt werden müssen.

Häufig verwendete Kriterien (nicht immer alle erfüllt):

- **nicht kompositionell:** Die Bedeutung einer Kollokation ist nicht aus den Bedeutungen ihrer Teile ableitbar.

  to kick the bucket

- **nicht austauschbar:** Teile der Kollokation können nicht durch semantisch äquivalente Ausdrücke ersetzt werden.

  to kick the bin

- **nicht modifizierbar:**

  to kick two buckets

- **nicht wörtlich übersetzbar:**

  to kick the bucket – *den Eimer treten

Indiz für Kollokationen: Eine Kollokation ist häufiger als aufgrund der Einzelwort-Häufigkeiten zu erwarten wäre.

Gründe für die Häufigkeit von Wortpaaren:

- Die Einzelwörter sind häufig.
- Es gibt syntaktische oder semantische Abhängigkeiten.
- Sie bilden eine Kollokation.

Bei der Kollokationsextraktion suchen wir Wortpaare, die häufiger sind, als auf Grund der Einzelwort-Häufigkeiten zu erwarten wäre.

Dazu wenden wir einen **statistischen Test** an.

Binomialtest

Beispiel: Ist das Wortpaar new companies signifikant häufiger als erwartet?

Daten: In einem Korpus mit $n=14{,}307{,}668$ Wörtern, taucht new $f_{new}=15{,}828$ Mal auf, companies $f_{companies}=4{,}675$ Mal und new companies 8 Mal.

Nullhypothese: Die gemeinsame Wahrscheinlichkeit p von new und companies ist nicht höher als das Produkt ihrer Randverteilungen, also maximal

$$p = \tilde{p}_{new}\tilde{p}_{companies} = \frac{f_{new}}{n}\frac{f_{companies}}{n} = \frac{15828}{14307668}\frac{4675}{14307668} = 3.615 \cdot 10^{-7}$$

Die Wahrscheinlichkeiten $\tilde{p}_{new}$ und $\tilde{p}_{companies}$ wurden hier mit relativen Häufigkeiten aus der Stichprobe geschätzt:

$$\tilde{p}_{new} = \frac{f_{new}}{n} \qquad \tilde{p}_{companies} = \frac{f_{companies}}{n}$$

Die Wahrscheinlichkeit, 8 oder mehr 1-Ereignisse (hier Wortpaare *new companies*) in n=14,307,668 Wiederholungen eines Bernoulliversuches mit der Wahrscheinlichkeit p zu bekommen, ist somit

$$b(\geq 8, n, p) = 1 - \sum_{i=0}^{7} b(i, n, p) \approx 0.15$$

$\Rightarrow$ Das Ergebnis ist **nicht** statistisch signifikant.

In der Praxis wenden wir statistische Tests bei der Kollokationsextraktion nicht an, um die **Signifikanz** zu berechnen, sondern nur um die Wortpaar-Kandidaten zu **ranken**.

Wir können dann bspw. die am höchsten gerankten Wortpaare manuell untersuchen, um echte Kollokationen zu extrahieren.

Die gleichzeitige Anwendung eines statistischen Testes auf viele Wortpaare ist auch gar nicht zulässig:

Wenn man bspw. bei 1000 Wortpaaren untersucht, ob sie signifikant häufiger auftreten, würde man ja erwarten, $1000 \cdot 0.05 = 50$ "signifikante" Wortpaare zu finden, auch wenn die Nullhypothese für alle 1000 Wortpaare stimmt.

Biased coin example

1. You toss a coin 80 times and get head 10 times and tail 70 times. Is the coin bias? and bias toward head or tail? Prove this using a binomial test.

Pick significant value = 0.05

H_0 : P(head) = 0.5

H_Alternative: P(head) < 0.5 (we suspect that the coin is biased toward tail)

compute P_binomial(X<= 10)

Beispiel: Ist das Wortpaar new companies signifikant häufiger als erwartet?

Daten: In einem Korpus mit n=14,307,668 Wörtern, taucht new f_{new}=15,828 Mal auf, companies $f_{companies}$=4,675 Mal und new companies 8 Mal.

Nullhypothese: Die gemeinsame Wahrscheinlichkeit p von new und companies ist nicht höher als das Produkt ihrer Randverteilungen, also maximal

$$p = \tilde{p}_{new}\tilde{p}_{companies} = \frac{f_{new}}{n}\frac{f_{companies}}{n} = \frac{15828}{14307668}\frac{4675}{14307668} = 3.615 \cdot 10^{-7}$$

Die Wahrscheinlichkeiten $\tilde{p}_{new}$ und $\tilde{p}_{companies}$ wurden hier mit relativen Häufigkeiten aus der Stichprobe geschätzt:

$$\tilde{p}_{new} = \frac{f_{new}}{n} \qquad \tilde{p}_{companies} = \frac{f_{companies}}{n}$$

Kollokation

You observe word pairs in a corpus of 14,307,668 bigrams and found „new company" 8 times. Do „new" and „company" have some dependency?

First, formulate this situation as a Bernoulli experiment
Experiment: observe the pairs (word1, word2) in a corpus
Sample space: { (new, company) (..., company) (new, ...) (the, dog), ... }

X=1 : events where we get „new company" = { (new, company)}
X=0 : events where we don't get „ new company" = { (..., company) (new, ...) (the, dog), ..}

- What we want to prove is „new" and „company" have dependency. (This is H_Alternative)

- So we set H_0 to be the opposite of what we believe (want to prove) and also define P(X=1) for H_0.
  - H_0: „new" and „company" are not dependent
  - P(X=1) is the same as p(new, company)
- „new" and „company" are dependent, when they occur more often than what is expected under H_0
  - So, to prove that our assumption is true (has significant evidence to reject H_0) we have to show that what we observed (8 times „new company") is considered as higher than expected under H_0.

First, find out what p(new, company) for H_0 is.

H_0 believes that the words are not dependent, which means that the joint probability p(new, company) must be the same as the p(new) * p(company)


Statistische Unabhängigkeit

Unabhängigkeit: Die Zufallsvariablen X und Y sind statistisch unabhängig, falls für alle x und y gilt:

$$p(x, y) = p_X(x) p_Y(y)$$

The frequency of new and company are given, so we can estimate p(new)*p(company) using the relative frequency.

Here we get p(new)*p(company) = 0.0000003615
Since p(new)*p(company) must be the same as p(new, company),
we can say that p(new, company) = 0.0000003615
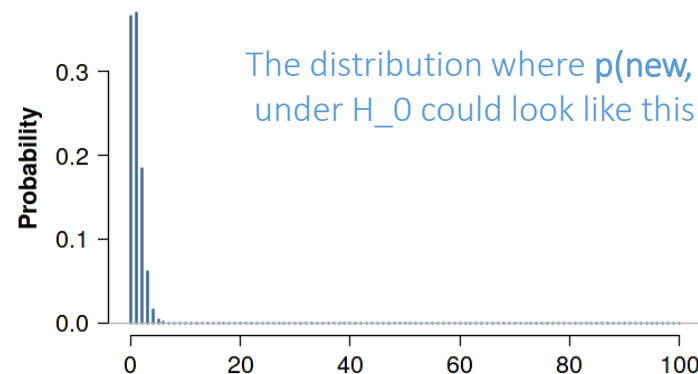
The distribution where p(new, company) = 0.0000003615 under H_0 could look like this

... 14,307,668

Beispiel: Ist das Wortpaar new companies signifikant häufiger als erwartet?

Daten: In einem Korpus mit $n=14{,}307{,}668$ Wörtern, taucht new $f_{new}=15{,}828$ Mal auf, companies $f_{companies}=4{,}675$ Mal und new companies 8 Mal.

Nullhypothese: Die gemeinsame Wahrscheinlichkeit p von new und companies ist nicht höher als das Produkt ihrer Randverteilungen, also maximal

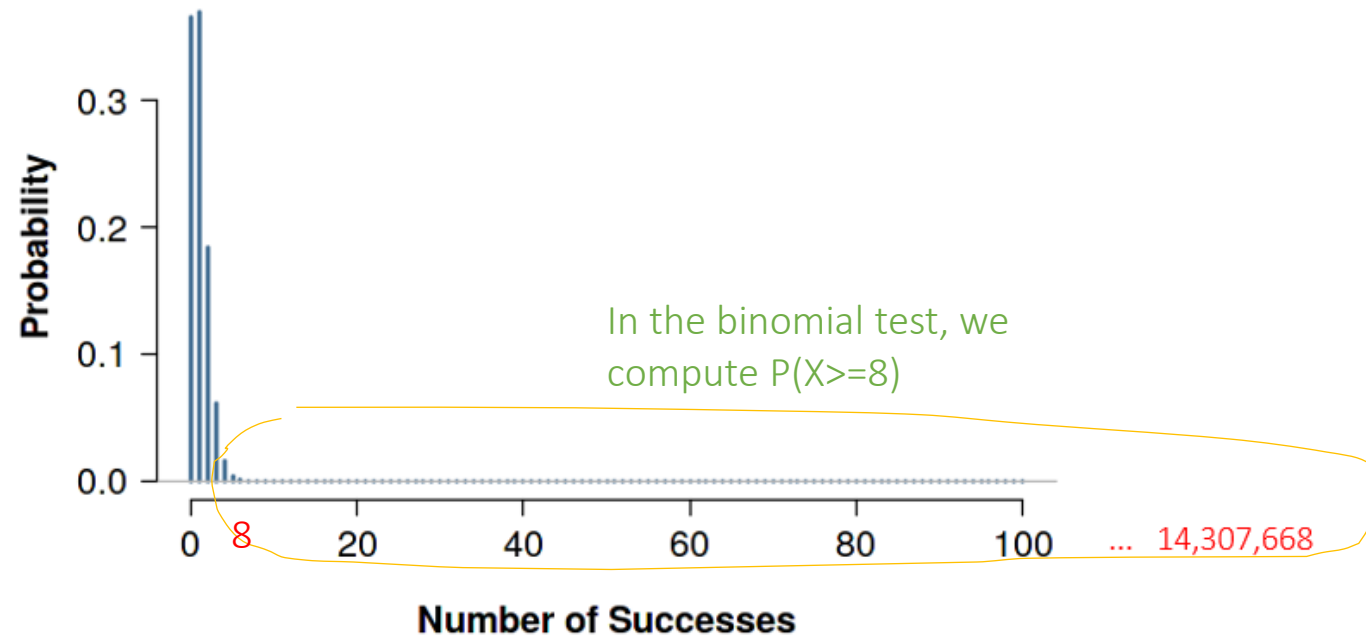$$p = \tilde{p}_{new}\,\tilde{p}_{companies} = \frac{f_{new}}{n}\,\frac{f_{companies}}{n} = \frac{15828}{14307668}\,\frac{4675}{14307668} = 3.615 \cdot 10^{-7}$$

Die Wahrscheinlichkeiten $\tilde{p}_{new}$ und $\tilde{p}_{companies}$ wurden hier mit relativen Häufigkeiten aus der Stichprobe geschätzt:

$$\tilde{p}_{new} = \frac{f_{new}}{n} \qquad \tilde{p}_{companies} = \frac{f_{companies}}{n}$$

Next, we have to prove that 8 occurences of „new company" is higher than what is expected in H_0

To do that, we can **not** simply compare n*p under H_0 and the real observation 8 but we have to use the binomial test (or other hypothesis tests). The test can tell us whether the result is **statistically significant** or not.

We can say that the freq 8 is higher than expected, when $P(X{>=}8)$ is a very low value. This is because low $P(X{>=}8)$ indicates that the freq 8 (and more) is not a common event to occur under H_0.

Now after computing $P(X{>=}8)$, we have to see if it is low enough that we can reject H_0. We compare $P(X{>=}8)$ with the significant value that we picked, for example, 0.05. If $P(X{>=}8)$ is lower than 0.05, then we have a reason (enough evidence) to reject H_0, otherwise, we can not reject H_0.

H_0 : p(new, company) = 0.0000003615

In the binomial test, we compute $P(X{>=}8)$

We have to compute $P(X{>=}8)$ not only $P(X{=}8)$ because we have to make sure that even when we see 9 or 10 or 11 occurrences of „new company" it will be still considered a rare event.

0   8   20   40   60   80   100   ... 14,307,668

Number of Successes

The freq that is higher than 8 would be even stronger support for our assumption (H_alternative: 8 is higher than what is expected under H_0). If we look at the graph, we can see that values like 9, 10, 11 are even far away from the mean in H_0.

the frequency of „new company"

Take a look at the distribution under H_0
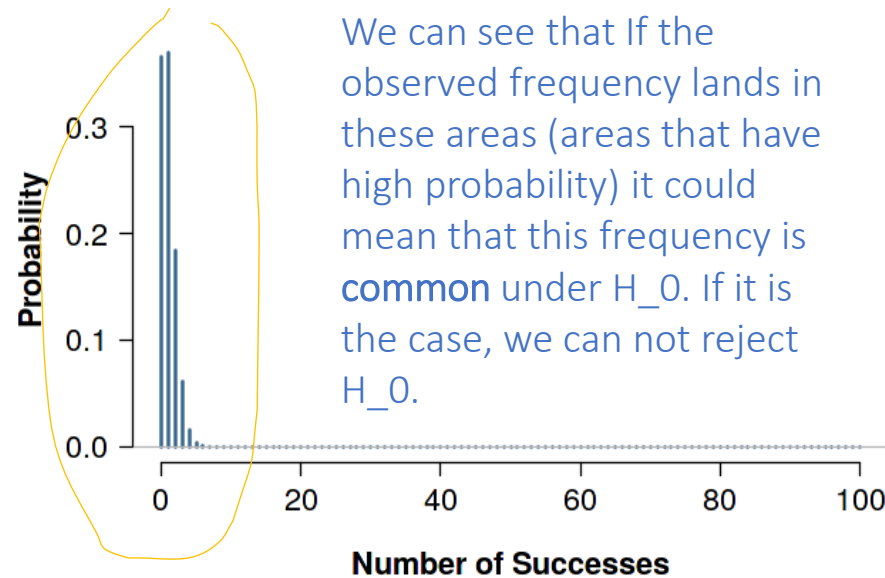
We can see that If the observed frequency lands in these areas (areas that have high probability) it could mean that this frequency is **common** under H_0. If it is the case, we can not reject H_0.
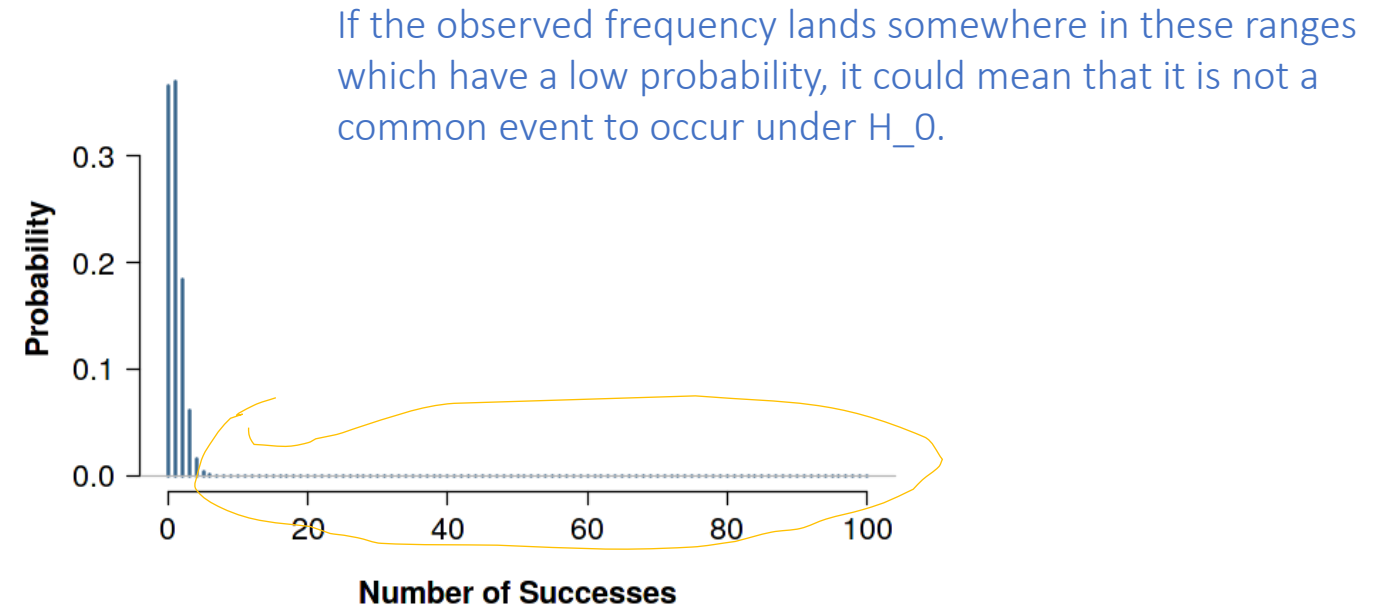
If the observed frequency lands somewhere in these ranges which have a low probability, it could mean that it is not a common event to occur under H_0.

What we do in the binomial test is measuring the probability of the entire range (from the observed freq until n, e.g. P(X>=8)) and see whether the probability is low enough (It is low enough when it is lower than our significant value) that we can reject H_0 and says that, what we observed is very unlikely to occur under H_0 and that is the reason for rejecting H_0.

Go back to our question

Kollokation

You observe word pairs in a corpus of 14,307,668 bigrams and found „new company" 8 times. Do „new" and „company" have some dependency?

First, formulate this situation as a Bernoulli experiment
Experiment: observe the pairs (word1, word2) in a corpus
Sample space: { (new, company) (..., company) (new, ...) (the, dog), ... }

X=1 : events where we get „new company" = { (new, company)}
X=0 : events where we don't get „ new company" = { (..., company) (new, ...) (the, dog), ..}

- What we want to prove is „new" and „company" have dependency.

- So we set H_0 to be the opposite of what we believe (want to prove) and also define P(X=1) for H_0.
- P(X=0) is P(new, company)
- „new" and „company" are dependent, when they occur more often than what is expected in H_0
- So, to prove that our assumption is true (has a sinificant evident to reject H_0) we have to show that what we observed (8 times „new company") is considered as higher than expected under H_0.

We have now
H_0: p(new, company) = 0.0000003615

Then we have to compute P(X>=8)

Die Wahrscheinlichkeit, 8 oder mehr 1-Ereignisse (hier Wortpaare *new companies*) in n=14,307,668 Wiederholungen eines Bernoulliversuches mit der Wahrscheinlichkeit *p* zu bekommen, ist somit

$$b(\geq 8, n, p) = 1 - \sum_{i=0}^{7} b(i, n, p) \approx 0.15$$

We see that P(X>=8) is around 0.15

If we pick a significant value of 0.05, this test result will be considered as "not significant", meaning it is not lower than 0.05. So we can not reject the null hypothesis.

Question from a student

ist unsere null-hypothese, dass sie nicht abhängig sind, weil sie so selten zusammen auftraten?

also wenn wir eine zahl z.B 500 und nicht 8 hätten

hätten wir dann als null-hypothese gesagt sie sind abhängig?

aber ich dachte die nullhypothese ist das was wir glauben?

weil wir hatten auch vorher gesagt dass die wahrscheinlichkeit 0,5 ist

als nullhypothese

For collocation, the probability under H_0 is not 0.5.
It is the case for the coin question.

H_0 is the opposite of what we believe/ want to prove. Normally, we would intentionally set H_0 to be something we want to reject. Something that, if we reject it, will support our assumption/belief.

We set H_0 to be „new company" are not dependent because we wanted to prove that they are dependent (this is our assumption H_A).

We did not set H_0 based on the fact that 8 in 14,307,668 is a very low frequency.

The problem definition is a bit different than in the biased coin case. The assumption in the collocation case is sort of already defined. It is already given that we want to know/prove that two words have some dependency. And this becomes our H_A.

In the biased coin case, we ask, is the coin biased and toward which side? Here, there are more ways to define the problem. That is why we have to look at the number of head we get and think about what we want to take as X=1 (success event) and so on.

Read more about collocations and hypothesis testing   https://nlp.stanford.edu/fsnlp/promo/colloc.pdf

5 *Collocations*

A COLLOCATION is an expression consisting of two or more words that correspond to some conventional way of saying things. Or in the words of Firth (1957: 181): "Collocations of a given word are statements of the habitual or customary places of that word." Collocations include noun phrases like *strong tea* and *weapons of mass destruction*, phrasal verbs like *to make up*, and other stock phrases like *the rich and powerful*. Particularly interesting are the subtle and not-easily-explainable patterns of word usage that native speakers all know: why we say *a stiff breeze* but not *??a stiff wind* (while either *a strong breeze* or *a strong wind* is okay), or why we speak of *broad daylight* (but not *?bright daylight* or *??narrow darkness*).

COMPOSITIONALITY   Collocations are characterized by limited *compositionality*. We call a natural language expression compositional if the meaning of the expression can be predicted from the meaning of the parts. Collocations are not fully compositional in that there is usually an element of meaning added to the combination. In the case of *strong tea*, *strong* has acquired the meaning *rich in some active agent* which is closely related, but slightly different from the basic sense *having great physical strength*. Idioms are the most extreme examples of non-compositionality. Idioms like *to kick the bucket* or *to hear it through the grapevine* only have an indirect historical relationship to the meanings of the parts of the expression. We are not talking about buckets

some related videos

7.1a Hypothesis Testing - Binomial
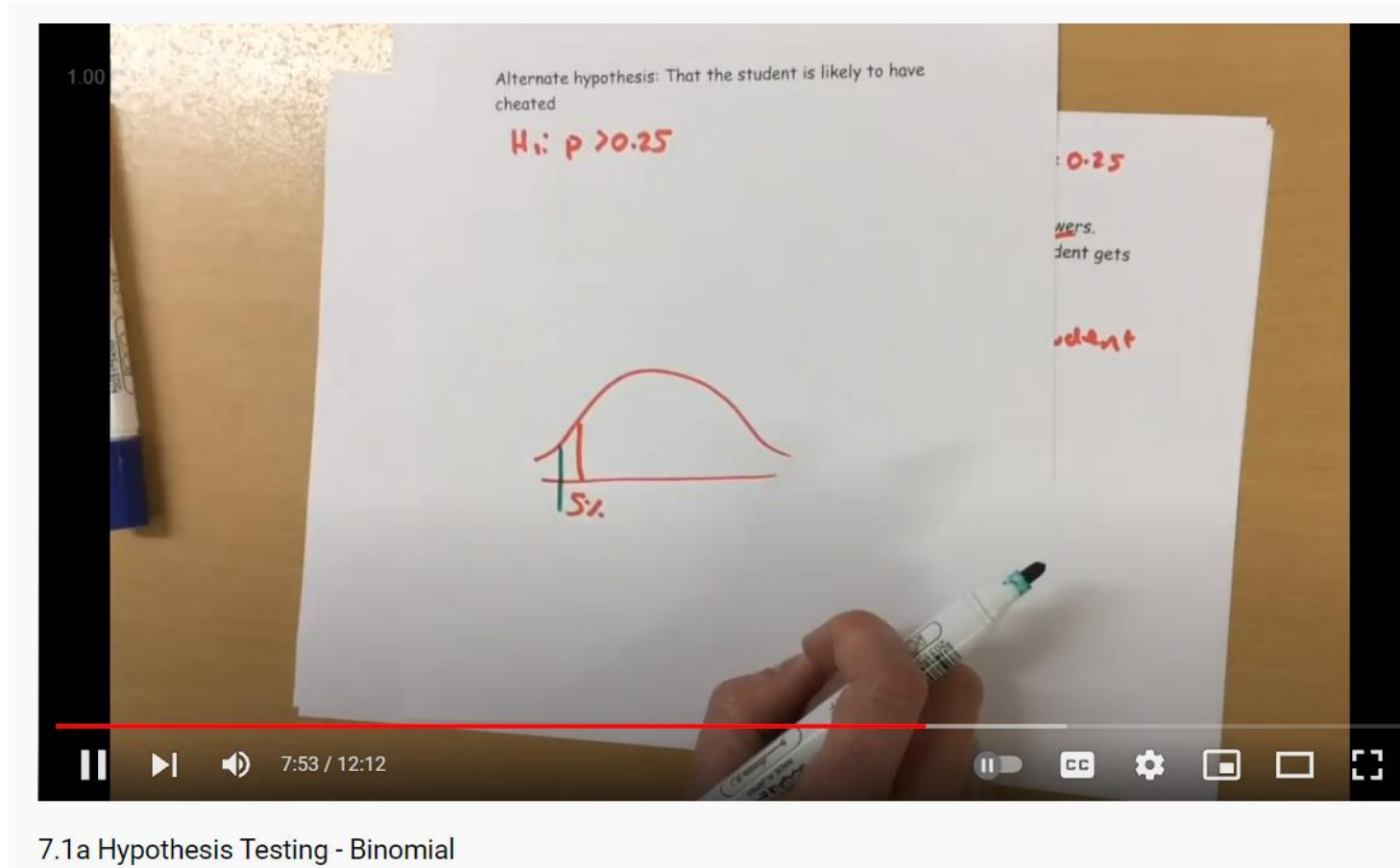
https://www.youtube.com/watch?v=_QuwqajyK0o&ab_channel=JoeBirch
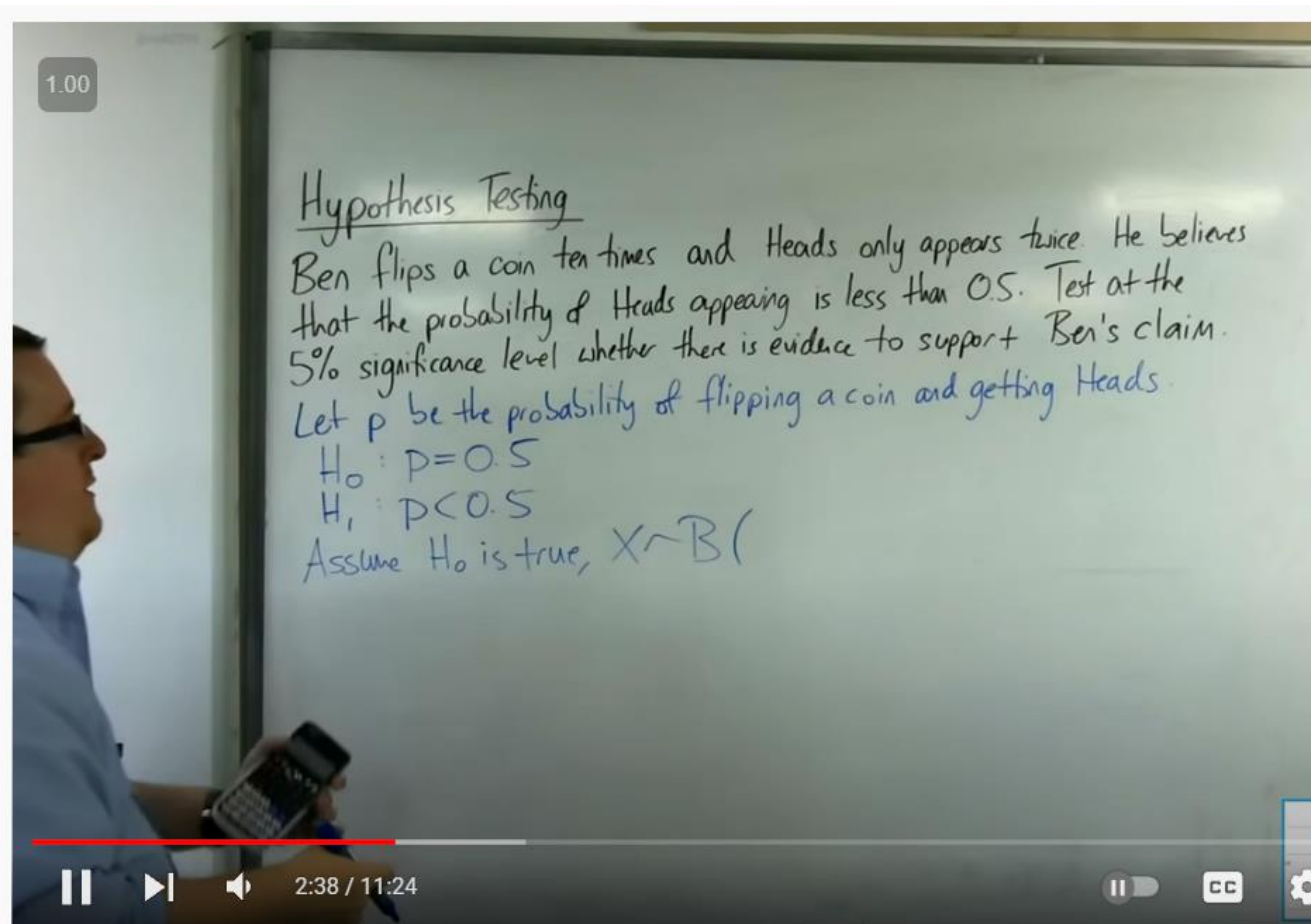
32

A-Level Maths: O2-01 [Binomial Hypothesis Testing: Less Than Example 1]

https://www.youtube.com/watch?v=TKRdk3NYkz8&t=602s&ab_channel=TLMaths

Kettenregel

Eine gemeinsame Wahrscheinlichkeit kann in ein Produkt bedingter Wahrscheinlichkeiten umgewandelt werden.

$$
\begin{aligned}
p(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) &= p(x_1, x_2, ..., x_n) \\
&= p(x_1)p(x_2|x_1)...p(x_n|x_1, ..., x_{n-1}) \\
&= \prod_{i=1}^{n} p(x_i|x_1, ..., x_{i-1})
\end{aligned}
$$

Übung

$p(W_1{=}Es, W_2{=}gibt, W_3{=}ein) =$

$p(Es, gibt, ein) =$

Kettenregel

Eine gemeinsame Wahrscheinlichkeit kann in ein Produkt bedingter Wahrscheinlichkeiten umgewandelt werden.

$$p(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = p(x_1, x_2, ..., x_n)$$
$$= p(x_1)p(x_2|x_1)...p(x_n|x_1, ..., x_{n-1})$$
$$= \prod_{i=1}^{n} p(x_i|x_1, ..., x_{i-1})$$

Übung

$$p(W_1=Es, W_2=gibt, W_3=ein) = \quad p(W_1=Es)\, p(W_2=gibt|W_1=Es)\, p(W_3=ein|W_1=Es, W_2=gibt)$$

$$p(Es, gibt, ein) = \quad p(Es)\, p(gibt|Es)\, p(ein|Es, gibt)$$

Theorem von Bayes

erlaubt es, eine bedingte Wahrscheinlichkeit "umzudrehen"

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

$p(W_2{=}York|W_1{=}New) =$

Theorem von Bayes

erlaubt es, eine bedingte Wahrscheinlichkeit "umzudrehen"

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

$$p(W_2{=}York|W_1{=}New) = \quad p(W_1{=}New|W_2{=}York)p(W_2{=}York)/p(W_1{=}New)$$

Informationsgehalt: $I(x) = -\log_2 p(x)$

- Der Informationsgehalt eines Ereignisses gibt an, wie informativ dieses Ereignis ist.
    - also seen as the level of surprise
    - high Info content -> high surprise, low prob

- Den Info-gehalt berechnet man für ein Ereignis
- x ist ein Wert einer Zufallsvariable. Es ist auch ein Ereignis.

Bsp.
Beim Münzenwurf haben wir Ergebnisraum = {Head, Tail}
- {Head} is das Ereignis, wo wir head bekommen
- {Tail} is das Ereignis, wo wir tail bekommen
- Was ist I ({head}) ?
    - Wir müssen p({Kopf}) kennen, was 0,5 ist
    - Dann berechne – log 0.5
                    = 1 bit

For calculating log

https://www.omnicalculator.com/math/log

https://www.rapidtables.com/calc/math/Log_Calculator.html

In information theory, the **information content**, **self-information**, **surprisal**, or **Shannon information** is a basic quantity derived from the probability of a particular event occurring from a random variable. It can be thought of as an alternative way of expressing probability, much like odds or log-odds, but which has particular mathematical advantages in the setting of information theory.

The Shannon information can be interpreted as quantifying the level of "surprise" of a particular outcome. As it is such a basic quantity, it also appears in several other settings, such as the length of a message needed to transmit the event given an optimal source coding of the random variable.

$\Rightarrow$ Je unwahrscheinlicher ein Ereignis ist, desto informativer ist es.

Informationsgehalt: $I(x) = -\log_2 p(x)$

Beispiel: Der Informationsgehalt des Ergebnisses eines Münzwurfes beträgt
$-\log_2 0.5 = 1\,\text{Bit}$

Bit ist die Maßeinheit der Information.

```
[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]

p= 0.1 , info content = 3.32 bit
p= 0.2 , info content = 2.32 bit
p= 0.3 , info content = 1.74 bit
p= 0.4 , info content = 1.32 bit
p= 0.5 , info content = 1.0 bit
p= 0.6 , info content = 0.74 bit
p= 0.7 , info content = 0.51 bit
p= 0.8 , info content = 0.32 bit
p= 0.9 , info content = 0.15 bit
p= 1.0 , info content = -0.0 bit
```

Entropie

Die **Entropie** misst, wieviel Information ein Zufallsereignis im Mittel enthält.

Entropie einer Zufallsvariablen X mit der Wahrscheinlichkeitsverteilung $p(x)$:

$$H(X) = H(p) = -\sum_{x \in \Omega_X} p(x) \log_2 p(x) = E\left(\log_2 \frac{1}{p(x)}\right)$$

Bsp. Beim Münzenwurf hat X 2 Werte
X=1 entpricht {head} mit p(head) = 0.5
X=0 entsptricht {tail} mit p(tail) = 0.5

Was ist H(X) ?

H(X) = - [p(1) log p(1) + p(2) log p(2)]
    = - [0.5 log 0.5+ 0.5 log 0.5]

Bsp. Der Würfel hat 6 mögliche Ausgänge (Ergebnisse)
Wie viele Informationen sind in diesem Zufallsexperiment enthalten?

X=1 entpricht {1}
X=2 entpricht {2}
X=3 entpricht {3}   p(x) = 1/6
X=4 entpricht {4}
X=5 entpricht {5}
X=6 entpricht {6}

H(X) = ?
    = - [p(1)log p(1) + p(2) log p(2) + ... + p(6)logp(6)]

Die Entropie beim Wurf eines Würfels beträgt
$-6 \cdot 1/6 \cdot \log_2 1/6 = \log_2 6 = 2,58 Bit$

Relationship to entropy [edit]

The entropy is the expected value of the information content of the discrete random variable, with expectation taken over the discrete values it takes. Sometimes, the entropy itself is called the "self-information" of the random variable, possibly because the entropy satisfies $\mathrm{H}(X) = \mathrm{I}(X;X)$, where $\mathrm{I}(X;X)$ is the mutual information of X with itself.[7]

For continuous random variables the corresponding concept is differential entropy.

Gemeinsame Entropie

```
result = 0
for x in X:
    for y in Y:
        result += p(x,y) * log p(x, y)
```

Die **gemeinsame Entropie** zweier Zufallsvariablen ist wie folgt definiert

$$H(X, Y) = - \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p(x, y) \; log_2 \; p(x, y)$$

Beispiel:

X	Y	$p(x,y)$	$p(y\|x)$
0	0	2/6	2/3
0	1	1/6	1/3
1	0	1/6	1/3
1	1	2/6	2/3

H(X, Y) = - [2/6 log 2/6 + 1/6 log 1/6 + 1/6 log 1/6 + 2/6 log 2/6   ]

What is X and Y?

Die **gemeinsame Entropie** zweier Zufallsvariablen ist wie folgt definiert

$$H(X, Y) = - \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p(x, y) \, log_2 \, p(x, y)$$

X, Y are random variables defined on the **same** sample space (S)

Examples
- **roll 1 die** ; S = {1,2,3,4,5,6}
    - X
        - X=1 ; {1,2}
        - X=0 ; {3, 4,5,6}
    - Y
        - Y=1 ; {2,4,6}
        - Y=0 ; {1,3,5}
- **toss 2 coins;** S= {(head, tail), (head, head), (tail, head), (tail, tail) }
    - X: result on coin 1
        - X=head ; {(head, tail), (head, head)}
        - X=tail ; {(tail, tail) , (tail, head)}
    - Y: result on coin 2
        - Y= head ; { (head, head) , (tail, head) }
        - Y=tail; { (tail, tail) , (head, tail) }

- roll 1 die ; S= {1,2,3,4,5,6}
    - X
        - X=1 ; {1,2}
        - X=0 ; {3, 4,5,6}
    - Y
        - Y=1 ; {1}
        - Y=2 ; {2}
        - Y=3 ; {3}
        - Y=4 ; {4}
        - Y=5 ; {5}
        - Y=6 ; {6}

Die **gemeinsame Entropie** zweier Zufallsvariablen ist wie folgt definiert

$$H(X, Y) = -\sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p(x, y) \log_2 p(x, y)$$

What is this?

- **roll 1 die** ; S = {1,2,3,4,5,6}
  - X
    - X=1 ; {1,2}
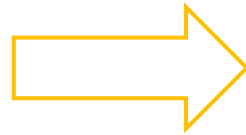   - X=0 ; {3, 4,5,6}
  - Y
    - Y=1 ; {2,4,6}
    - Y=0 ; {1,3,5}

Compute all combinations of p(x,y)

X	Y	$p(x, y)$
0	0	2/6
0	1	1/6
1	0	1/6
1	1	2/6

Another example of how to compute p(x,y)
This is the experiment where we roll 2 dice.

Define 2 random variables W1, W2

All possible joint prob p(x,y)

sample space has 36 outcomes

w2					
1	**2**	**3**	**4**	**5**	**6**

w1						
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

```
p(W1=1, W2=1)          p(W1=4, W2=1)    = P ( { (4,1) } ) = 1 /36
p(W1=1, W2=2)          p(W1=4, W2=2)
p(W1=1, W2=3)          p(W1=4, W2=3)
p(W1=1, W2=4)          p(W1=4, W2=4)
p(W1=1, W2=5)          p(W1=4, W2=5)
p(W1=1, W2=6)          p(W1=4, W2=6)


p(W1=2, W2=1)          p(W1=5, W2=1)
p(W1=2, W2=2)          p(W1=5, W2=2)
p(W1=2, W2=3)          p(W1=5, W2=3)
p(W1=2, W2=4)          p(W1=5, W2=4)
p(W1=2, W2=5)          p(W1=5, W2=5)
p(W1=2, W2=6)          p(W1=5, W2=6)


p(W1=3, W2=1)          p(W1=6, W2=1)
p(W1=3, W2=2)          p(W1=6, W2=2)
p(W1=3, W2=3)          p(W1=6, W2=3)
p(W1=3, W2=4)          p(W1=6, W2=4)
p(W1=3, W2=5)          p(W1=6, W2=5)
p(W1=3, W2=6)          p(W1=6, W2=6)
```

1/36

Die **gemeinsame Entropie** zweier Zufallsvariablen ist wie folgt definiert

$$H(X, Y) = - \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p(x, y) \, log_2 \, p(x, y)$$

Beispiel: Die Entropie $H(W_1, W_2)$ beim Wurf von 2 Würfeln beträgt

Die **gemeinsame Entropie** zweier Zufallsvariablen ist wie folgt definiert

$$H(X, Y) = - \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p(x, y) \, log_2 \, p(x, y)$$

Beispiel: Die Entropie $H(W_1, W_2)$ beim Wurf von 2 Würfeln beträgt
$$-36 \cdot 1/36 \cdot log_2 \, 1/36 = log_2 \, 36 = 5,17 \text{ Bit}$$

$W_1 =$ Augenzahl des 1. Würfels

$W_2 =$ Augenzahl des 2. Würfels

Joint entropy

In information theory, **joint entropy** is a measure of the uncertainty associated with a set of variables.[2]

$H(X, Y)$ represents the amount of randomness in both X and Y, or the number of bits required to describe both of them. We now explore

https://www.sciencedirect.com/topics/mathematics/joint-entropy

Bedingte Entropie

$$H(Y|X) = \sum_{x \in \Omega_X} p(x) H(Y|X = x)$$

$$= \sum_{x \in \Omega_X} p(x) \left[-\sum_{y \in \Omega_Y} p(y|x) \, log_2 \, p(y|x) \right]$$

$$= -\sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p(x,y) \, log_2 \, p(y|x)$$

Beispiel:

X	Y	$p(x,y)$	$p(y|x)$
0	0	2/6	2/3
0	1	1/6	1/3
1	0	1/6	1/3
1	1	2/6	2/3

Die bedingte Entropie $H(Y|X)$ beim Wurf eines Würfels mit
$X = 1$ falls Augenzahl > 3 sonst 0
$Y = 1$ falls Augenzahl gerade sonst 0
beträgt:
$-(2 \cdot 2/6 \, log_2 \, 2/3 + 2 \cdot 1/6 \, log_2 \, 1/3) = 0.92$

Conditional Entropy

- $H(Y|X) = 0$, if and only if, the value of Y is completely determined by the value of X

- $H(Y|X) = H(Y)$, if and only if, Y and X are independent random variables

  - X does not provide any information about Y

  - Entropy of Y given X is simply equal as asking for the Entropy of Y directly with no access to X

Relative Entropie

Die **Relative Entropie** (Kullback-Leibler-Abstand) zwischen zwei
Verteilungsfunktionen $p(x)$ und $q(x)$ ist wie folgt definiert:

$$D(p||q) = \sum_{x \in \Omega_X} p(x) \, log_2 \, \frac{p(x)}{q(x)}$$

Die relative Entropie ist nie negativ, nicht symmetrisch und 0 falls $p = q$.
Sie wird oft als Abstandsmaß für Wahrscheinlichkeitsverteilungen verwendet.

Die Begriffe **Kullback-Leibler-Divergenz** (kurz *KL-Divergenz*) und **Kullback-Leibler-Abstand** (auch *Kullback-Leibler-Entropie*
oder *Kullback-Leibler-Information*, nach Solomon Kullback und Richard Leibler; englisch *Information Gain*) bezeichnen ein Maß
für die Unterschiedlichkeit zweier Wahrscheinlichkeitsverteilungen. Typischerweise repräsentiert dabei eine der Verteilungen
empirische Beobachtungen oder eine präzise Wahrscheinlichkeitsverteilung, während die andere ein Modell oder eine
Approximation darstellt.

Die KL-Divergenz wird auch *relative Entropie* genannt, wobei der Begriff *relative Entropie* gelegentlich auch für die
Transinformation verwendet wird.
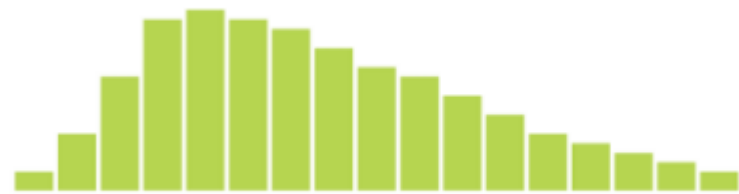
Positive Skew
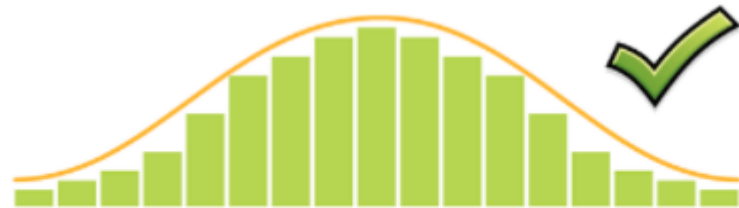
Normal Distribution
No Bias

- wie sehen die Verteilungsfunktionen p(x) und q(x) aus?

Example

You have a random variable X.
The distribution function p(x) might look like the first histogram.
q(x) is another distribution function for the same random variable x (the second histogram), but it assigns probability to the individual events/outcomes in a different way than p(x).

Relative Entropie

Die relative Entropie gibt an, wieviele Bits im Mittel verschenkt werden,
wenn Ereignisse mit einer Verteilung p mit einem Code kodiert werden,
der optimal für die Verteilung q ist.

Beispiel:

x	p(x)	Code(x)	q(x)
a	0.45	1	0.5
e	0.30	01	0.25
i	0.15	001	0.125
o	0.10	000	0.125

$$D(p\|q) = 0.45 \log_2 \frac{0.45}{0.5} + 0.30 \log_2 \frac{0.30}{0.25} + 0.15 \log_2 \frac{0.15}{0.125} + 0.10 \log_2 \frac{0.10}{0.125} \approx 0.01777 \text{ Bit}$$

Das ist nicht dasselbe wie

$$D(q\|p) = 0.5 \log_2 \frac{0.5}{0.45} + 0.25 \log_2 \frac{0.25}{0.30} + 0.125 \log_2 \frac{0.125}{0.15} + 0.125 \log_2 \frac{0.125}{0.10} \approx 0.01760 \text{ Bit}$$

$D(p\|q)$ wird unendlich, falls es ein x gibt mit $p(x) > 0$ und $q(x) = 0$,
d.h. falls es ein Symbol gibt, das nicht kodiert werden kann!

Crossentropie

Die **Crossentropie** zwischen zwei Verteilungen p und q

$$H(p, q) = -\sum_{x} p(x) \log_2 q(x)$$

$H(p,q) = - [\ p(a) \log q(a) + p(e) \log q(e) + p(i) \log q(i) + p(o) \log q(o)\ ]$

x	p(x)	q(x)
a	0.45	0.5
e	0.30	0.25
i	0.15	0.125
o	0.10	0.125

Cross-Entropy as a Loss Function

Cross-entropy is widely used as a loss function when optimizing classification models.

Two examples that you may encounter include the logistic regression algorithm (a linear classification algorithm), and artificial neural networks that can be used for classification tasks.

> ... using the cross-entropy error function instead of the sum-of-squares for a classification problem leads to faster training as well as improved generalization.

— Page 235, Pattern Recognition and Machine Learning, 2006.

Calculate Cross-Entropy Between Class Labels and Probabilities

The use of cross-entropy for classification often gives different specific names based on the number of classes, mirroring the name of the classification task; for example:

- **Binary Cross-Entropy**: Cross-entropy as a loss function for a binary classification task.
- **Categorical Cross-Entropy**: Cross-entropy as a loss function for a multi-class classification task.

Kreuzentropie

Die **Kreuzentropie** ist in der Informationstheorie und der mathematischen Statistik ein Maß für die Qualität eines Modells für eine Wahrscheinlichkeitsverteilung. Eine Minimierung der Kreuzentropie in Bezug auf die Modellparameter kommt einer Maximierung der Log-Likelihood-Funktion gleich.

Cross entropy

From Wikipedia, the free encyclopedia

In information theory, the **cross-entropy** between two probability distributions p and q over the same underlying set of events measures the average number of bits needed to identify an event drawn from the set if a coding scheme used for the set is optimized for an estimated probability distribution q, rather than the true distribution p.

Die **Crossentropie** eines Korpus $x_1^n = x_1 x_2 ... x_n$ wird folgendermaßen definiert:

x ist ein Wort

$$H(x_1^n, p) = -\frac{1}{n} \log_2 p(x_1^n)$$

p of this corpus

number of words in the corpus

Understanding perplexity, bits-per-character, and cross entropy

Language model

Consider an arbitrary language L. In this case, English will be utilized to simplify the arbitrary language. A language model assigns probabilities to sequences of arbitrary symbols such that the more likely a sequence $(w\_1, w\_2, ..., w\_n)$ is to exist in that language, the higher the probability. A symbol can be a character, a word, or a sub-word (e.g. the word 'going' can be divided into two sub-words: 'go' and 'ing'). Most language models estimate this probability as a product of each symbol's probability given its preceding symbols:

$$P_{(w_1, w_2, \ldots, w_n)} = p(w_1)p(w_2|w_1)p(w_3|w_1, w_2)...p(w_n|w_1, w_2, .., w_{n-1})$$

$$= \prod_{i=1}^{n} p(w_i|w_1, ..., w_{i-1}) \qquad (1)$$

```
S = Where are we going
```

Crossentropie

Die **Crossentropie** zwischen zwei Verteilungen p und q

$$H(p, q) = -\sum_x p(x) \, log_2 \, q(x)$$

$$= E_p(log_2 \frac{1}{q(x)})$$

$$= H(p) + D(p||q)$$

used as a loss function for classification task (machine learning)

Die **Crossentropie** eines Korpus $x_1^n = x_1 x_2 ... x_n$ wird folgendermaßen definiert:

$$H(x_1^n, p) = -\frac{1}{n} \, log_2 \, p(x_1^n)$$

used to measure the quality of a language model

Die **Perplexität** ist eng mit der Crossentropie verwandt:

$$perp(x_1^n, p) = 2^{H(x_1^n, p)} = p(x_1^n)^{-\frac{1}{n}}$$

$$I(x) = -\log_2 p(x)$$

we can measure the **info content** of an **event** using p(x)

$$H(X) = H(p) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Entropie von X

Mutual Information
$$H(X) - H(X|Y) = H(Y) - H(Y|X) =: I(X;Y)$$

X=1 is an event
p(X=1) = 0.3
X=1

p(X=2) = 0.7
X=2

X

$$H(X|Y) = -\sum_{x \in X}\sum_{y \in Y} p(x,y) \, \log_2 \, p(y|x)$$

bedingte Entropie von X, Y

need to know
p(x|y) and p(x,y) for every x, y

Y

Y=1

Y=2

X, Y are from the same sample space

gemeinsame Entropie
von X, Y

$$H(X,Y) = -\sum_{x \in X}\sum_{y \in Y} p(x,y) \, \log_2 \, p(x,y)$$

need to know
p(x,y) for every value x, y

derived
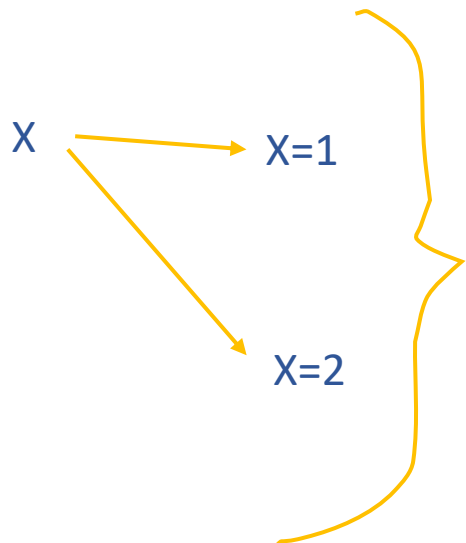
$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

Beziehung zw. gemeinsame und bedingte entropy

#summary

60

X → X=1

X=2

p(x)
p(1) = 0.5
p(2) = 0.5

q(x)
q(1) = 0.3
q(2) = 0.7

we can define 2 distributions p
and q for the same random
variable X

If we have p and q we can compute the
relative entropy and the cross-entropy

Die **Relative Entropie** (Kullback-Leibler-Abstand) zwischen zwei
Verteilungsfunktionen $p(x)$ und $q(x)$ ist wie folgt definiert:

$$D(p\|q) = \sum_{x \in \Omega_X} p(x) \ log_2 \ \frac{p(x)}{q(x)}$$

Die **Crossentropie** zwischen zwei Verteilungen p und q

$$
\begin{aligned}
H(p, q) &= -\sum_x p(x) \ log_2 \ q(x) \\
&= E_p(log_2 \frac{1}{q(x)}) \\
&= H(p) + D(p\|q)
\end{aligned}
$$

These are 3 ways for
computing H(p,q)