

- Die Vorlesungsfolien wurden geändert/verbessert, vergiss nicht, die neue Version herunterzuladen.

Fragen beantworten

Entropie

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

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

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

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

Beispiel

Y ist eine Zufallsvariable mit den folgenden Werten y und seine $p(y)$

$Y=1$ mit $p(1) = 0.4$

$Y=2$ mit $p(2) = 0.6$

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

- wie berechnet man die einzelne Entropie $H(Y)$?

Mutual Information

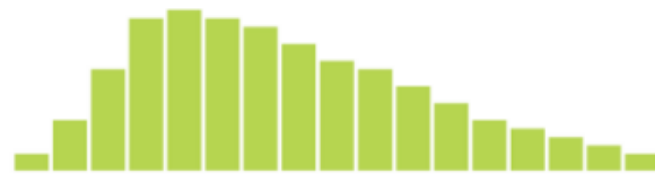
Die **Mutual Information** $I(X; Y)$ ist ein nicht negatives, symmetrisches Maß der gemeinsamen Information zweier Zufallsvariablen.

$$I(X; Y) = \sum_{x,y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}$$

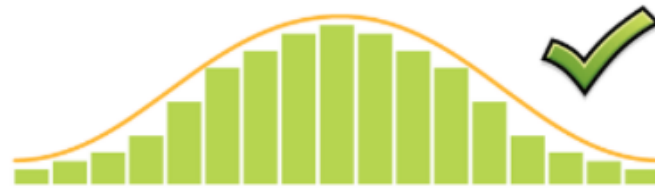
- wieso ist mutual information ein symmetrisches Maß?

Antwort

Denn man kann die Positionen der Argumente X und Y vertauschen und erhält trotzdem das gleiche Ergebnis



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 $p(x)$.

bedingte Wahrscheinlichkeit: $p(x|y) = \frac{p(x,y)}{p(y)}$

Kettenregel $p(x, y, z) = p(x)p(y|x)p(z|xy)$

Bayes'sches Theorem: $p(x|y) = \frac{p(y|x)p(x)}{p(y)}$

Unterschied zwischen $p(x|y)$ bedingte Wahrscheinlichkeit und $p(x|y)$ Bayes'sches Theorem?

Antwort: $p(x|y) = p(x,y)/p(y)$ ist die Definition der bedingten Wahrscheinlichkeit. Aus dieser Definition und der Kommutativität von $p(x,y) = p(y,x)$ lässt sich das Bayes'sche Theorem ableiten, wie in Übung 2 gezeigt wird.

Kollokationen

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

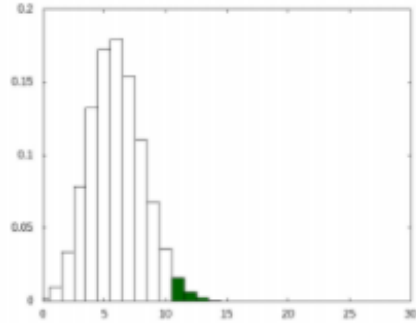
Häufig verwendete Kriterien (nicht immer alle erfüllt):

- **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

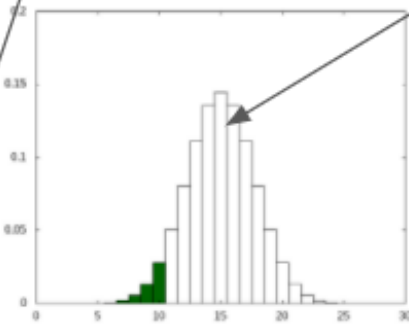
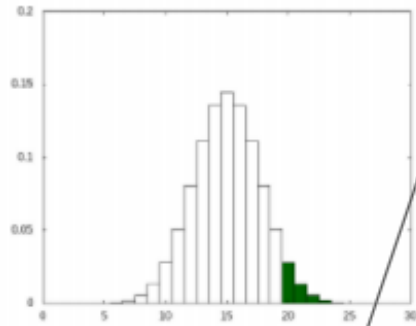
Indiz für Kollokationen: Eine Kollokation ist häufiger als aufgrund der Einzelwort-Häufigkeiten zu erwarten wäre.

- When two words occur frequently together, it does not always mean that they are a collocation or are interdependent in some way.
- It may simply be because both words occur frequently regardless of the other word.
- If these two words occur together more often than we would expect, it could be an indication that they are a collocation.
- We use a binomial test to find out if the frequency of the pair of words we observe in a corpus is significantly higher than the expected frequency.

Statistische Tests



Der grüne Bereich umfasst alle Ergebnisse, die statistisch signifikant sind (hier 10 oder mehr). Die Summe ihrer Wahrscheinlichkeiten ist maximal 0.05.



Falls $P(X \geq 10)$ höher als 0.05, dann können wir die Nullhypothese nicht ablehnen.

Eine Column ist die Wahrscheinlichkeitsmasse eines Wertes der Zufallsvariablen. In dieser Folie handelt es sich um eine **Binomialverteilung**. Wenn man ein Bernoulli-Experiment 29 Mal durchführt, dann hat die Zufallsvariable für die Binomialverteilung $29 + 1 = 30$ Werte (null Mal, zwei Mal, drei Mal, ...). Die Binomialverteilung gibt an, wie groß die Wahrscheinlichkeit ist, dass das Ereignis 1 (aus dem Bernoulli-Experiment) n-mal auftritt. Die Zufallsvariable des Bernoulli-Experiments hat zwei Werte 0 und 1. Für die Binomialverteilung ist nur der Wert 1 interessant.

Hier wird **10 oder mehr** nur als Beispiel verwendet.

Bei einem Hypothesentest wird man z.B. $P(X \geq 10)$ oder $P(X \leq 10)$ berechnen, wobei P eine Binomialverteilung ist. Der grüne Bereich repräsentiert also $P(X \geq 10)$ oder $P(X \leq 10)$.

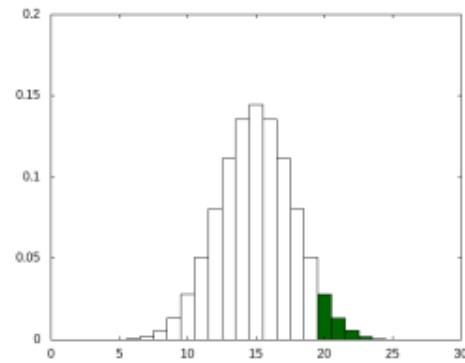
All dies wird getan, um die Werte des Tests zu berechnen. Dieser Wert wird dann mit dem Signifikanzniveau (z.B. 0,05) verglichen, um zu entscheiden, ob die Nullhypothese abgelehnt werden soll oder nicht.

- Seite 53: wofür steht ein Column? Wieso ist hier die Anzahl der Ergebnisse im grünen Bereich = 10 oder mehr?

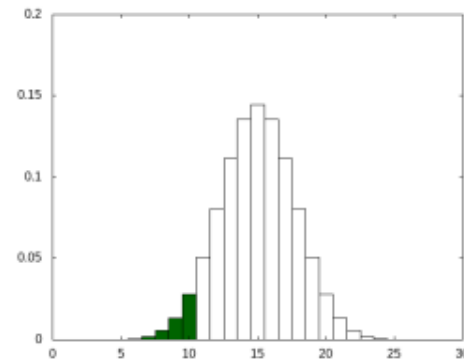
Diese Folie wurde bearbeitet, um Unklarheiten zu beseitigen. Die neue Version ist auf der nächsten Seite zu finden.

Statistische Tests

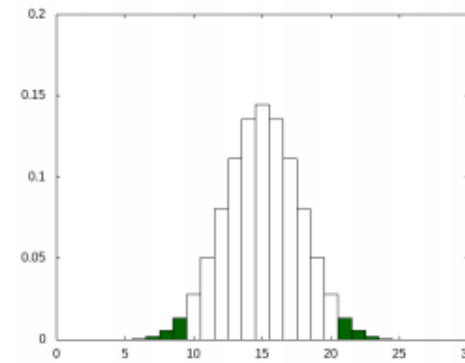
Es gibt drei Varianten von statistischen Tests:



rechtsseitiger Test



linksseitiger Test



beidseitiger Test

Beim beidseitigen Test interessieren wir uns für Abweichungen in beide Richtungen. Die Gesamtwahrscheinlichkeit muss hier auf jeder Seite kleiner als $0.05/2$ sein.

Statistische Tests

$$p(w_1, w_2) = p(w_1) \cdot p(w_2)$$

Nullhypothese: Die Einzelwörter sind **statistisch unabhängig**.

Unter der Annahme der Nullhypothese **erwarten** wir ungefähr $n \cdot p$ Vorkommen des Wortpaares (Korpusgröße n , Wortpaar-Wk. p)

Wenn die beobachtete Häufigkeit weit von der erwarteten Häufigkeit entfernt ist, **weisen wir die Nullhypothese zurück**. Das beobachtete Ergebnis ist zu unwahrscheinlich, um mit der Nullhypothese erklärt werden zu können.

Da die Wahrscheinlichkeit des beobachteten Ergebnisses zwar sehr klein aber nicht 0 ist, machen wir jedoch möglicherweise einen **Fehler**, wenn wir die Nullhypothese verwerfen.

Wir sind bereit, in maximal **5%** der Fälle, in denen wir die Nullhypothese verwerfen, einen Fehler zu machen.

Die Gesamtwahrscheinlichkeit aller Ergebnisse, bei denen wir die Nullhypothese verwerfen, darf also maximal 0.05 sein.

Frage: Was für ein Fehler ist gemeint? z.B x ist tatsächlich spam aber wir sagen x ist ham vor?

Antwort: Fehler = Die Wahrscheinlichkeit, dass die Nullhypothese zutrifft. D.h. obwohl wir aufgrund des Testergebnisses glauben, dass die einzelnen Wörter nicht unabhängig sind. Wir sind aber nicht 100% sicher, sondern nur 95% sicher.

Frage: "...weisen wir die Nullhypothese zurück" heisst wir nehmen an die Einzelwörter sind nicht mehr statistisch unabhängig?

Antwort: Ja, richtig.

Hier geht es darum, herauszufinden, ob zwei einzelne Wörter eines Wortpaares statistisch voneinander abhängig sind.

Zu diesem Zweck führen wir einen statistischen Test durch. Wir setzen die Nullhypothese auf das Gegenteil von dem, was wir annehmen, und führen den Test durch.

"...weisen wir die Nullhypothese zurück" bedeutet, dass für das untersuchte Wortpaar genügend Beweise vorliegen, um zu zeigen, dass die einzelnen Wörter des Paares statistisch nicht unabhängig sind. Mit anderen Worten, wir glauben jetzt, dass die einzelnen Wörter des Paares nicht unabhängig sind.

Das glauben aber wir mit 95%iger Sicherheit.

D.h. 5% ist die Chance, dass das nicht stimmt. Diese 5% sind ein Signifikanzniveau, das wir gewählt haben.

Wenn wir sicherer sein wollen, dann nehmen wir ein niedriges Signifikanzniveau (Fehlerwahrscheinlichkeit).

Diese Folie wurde bearbeitet, um Unklarheiten zu beseitigen. Die neue Version ist auf der nächsten Seite zu finden.

Statistische Tests

Ähnlich wie bei einem Widerspruchsbeweis machen wir eine Annahme (Nullhypothese), die wir widerlegen wollen, und zeigen dann, dass die beobachteten Stichprobenresultate der Annahme widersprechen.

Nullhypothese: Die Wahrscheinlichkeit von “Kopf” ist 0.5.

Unter der Annahme der Nullhypothese **erwarten** wir ungefähr $n \cdot p = 30 \cdot 0.5 = 15$ Mal “Kopf” zu sehen (Stichprobengröße n , Kopf-Wahrscheinlichkeit p)

Wenn die beobachtete Häufigkeit viel größer als die erwartete Häufigkeit ist, nehmen wir an, dass **die Nullhypothese falsch ist**, weil das beobachtete Resultat zu unwahrscheinlich ist, um mit der Nullhypothese erklärt werden zu können.

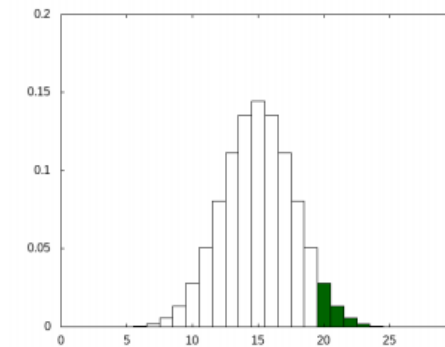
In unserem Fall würden wir dann schlussfolgern, dass die Münze gezinkt ist.

Statistische Tests

Da auch sehr unwahrscheinliche Resultate keine Wahrscheinlichkeit von 0 haben, machen wir möglicherweise einen **Fehler**, wenn wir die Nullhypothese verwerfen.

Wir sind bereit zu akzeptieren, dass in maximal **5%** der Fälle, in denen wir die Nullhypothese verwerfen, die Nullhypothese tatsächlich doch korrekt war. Die Gesamtwahrscheinlichkeit aller Resultate, bei denen wir die Nullhypothese verwerfen, darf daher maximal 0.05 sein.

Dann sprechen wir von einem **signifikanten Ergebnis**.



Die Grafik zeigt die Wahrscheinlichkeiten der möglichen Stichproben-Resultate unter Annahme der Nullhypothese.

Der grüne Bereich umfasst alle Resultate, die statistisch signifikant sind.

Die Summe ihrer Wahrscheinlichkeiten ist maximal 0.05.

Diese Folie wurde geändert.

Wahrscheinlichkeitsverteilung

Wahrscheinlichkeitsverteilung: Funktion, die jedem Ergebnis o einen Wert zwischen 0 und 1 zuweist, so dass

$$\sum_o p(o) = 1$$

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

Beispiel: Wahrscheinlichkeit des Ereignisses e , dass die Zahl der Augen beim Wurf eines Würfels gerade ist:

$$p(e) = p(2) + p(4) + p(6) = \frac{3}{6} = \frac{1}{2}$$

Das Problem war, dass p hier für Ereignis(e) und Ergebnis(o) verwendet wird.
Aber $p(e)$ und $p(o)$ sind nicht dieselbe Funktion.

Hier wurde der Ergebnisraum (ω) hinzugefügt, damit wir besser sehen können, woher o kommt.

Wahrscheinlichkeitsverteilung

Wahrscheinlichkeitsverteilung: Funktion, die jedem Ergebnis o einen Wert zwischen 0 und 1 zuweist, so dass gilt:

$$\sum_{o \in \Omega} p(o) = 1$$

Die Wahrscheinlichkeit $P(A)$ eines Ereignisses A ist die Summe der Wahrscheinlichkeiten der zugehörigen Ergebnisse.

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

Beispiel: Die Wahrscheinlichkeit des Ereignisses A , dass die Zahl der Augen beim Wurf eines Würfels gerade ist:

$$P(A) = p(2) + p(4) + p(6) = \frac{3}{6} = \frac{1}{2}$$

Anmerkung: $P(\cdot)$ ist keine Wahrscheinlichkeitsverteilung, da die Summe der Wahrscheinlichkeiten aller möglichen Ereignisse größer als 1 ist.

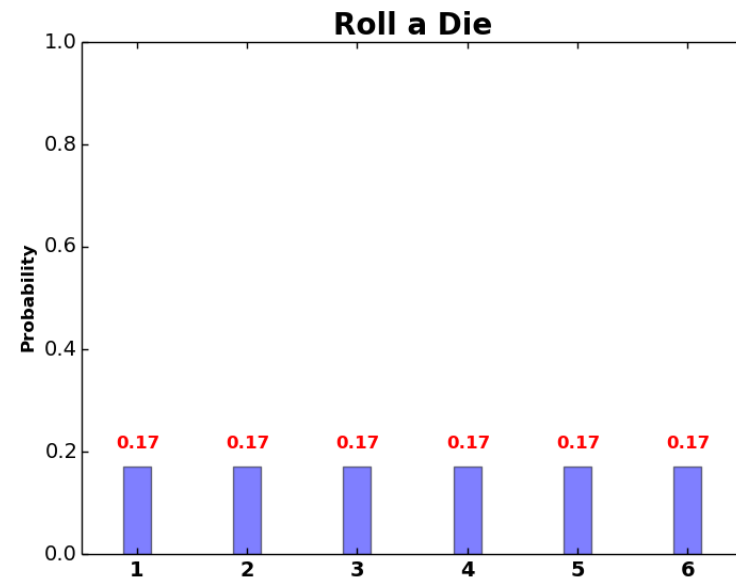
Hier wird P anstelle von p als Wahrscheinlichkeitsfunktion für Ereignisse verwendet.

More about probability distribution

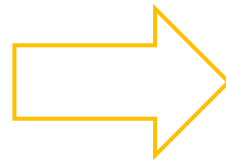
- Probability distribution is a general term that refers to a function P that maps sets of outcomes from a sample space to a probability.
- Different books define and use this term in different ways. In some books it is defined as a probability function for events (sets of outcomes) rather than for basic outcomes as in our lecture, so don't get confused if you read other sources.
- Regardless of the formal definition, an important property of a probability distribution is that $p(\text{sample space})$ must be 1.

We can also define a probability distribution for a random variable

- Suppose we have a random variable X and we want to assign a probability to each value x of X (e.g., $P(X=1) = 0.5$), we can define a probability distribution for this random variable. This is not the same probability distribution as the probability distribution for the events/outcomes.



probability distribution over a sample space



For example, we define a random variable X on this sample space as follows.

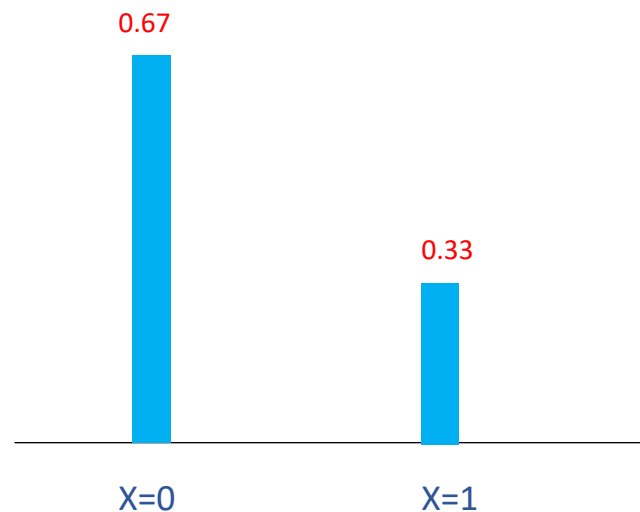
$X=1$ if $\{5,6\}$ and $X=0$ if $\{1,2,3,4\}$

So, $P(X=1) = P(\{5,6\}) = P(\{5\}) + P(\{6\}) \approx 0.17 + 0.17 \approx 0.33$

$P(X=0) = P(\{1\}) + P(\{2\}) + P(\{3\}) + P(\{4\})$

$\approx 0.17 + 0.17 + 0.17 + 0.17$

≈ 0.67



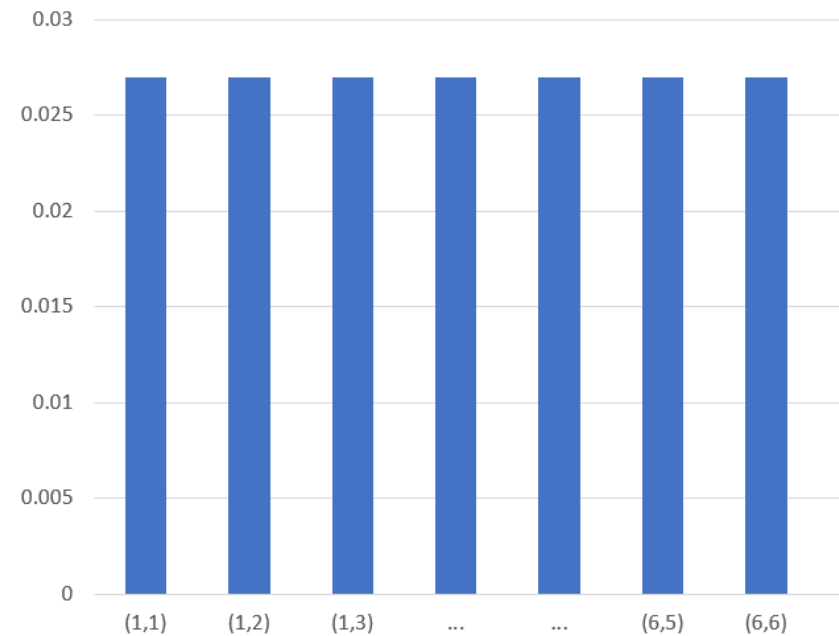
Example

Probability – Sample space for two dice (outcomes):

	1	2	3	4	5	6
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)

$$p = 1/36 = 0.027$$

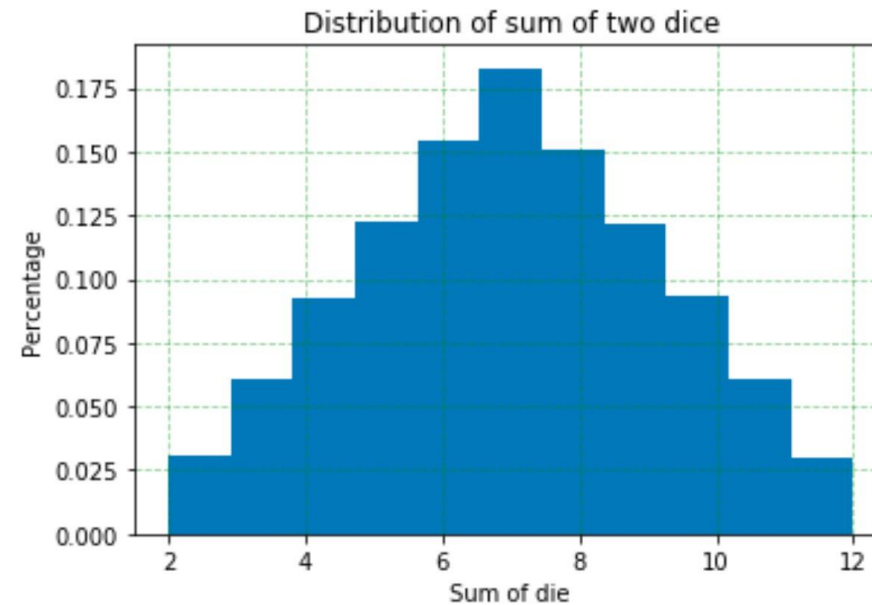
roll 2 dice



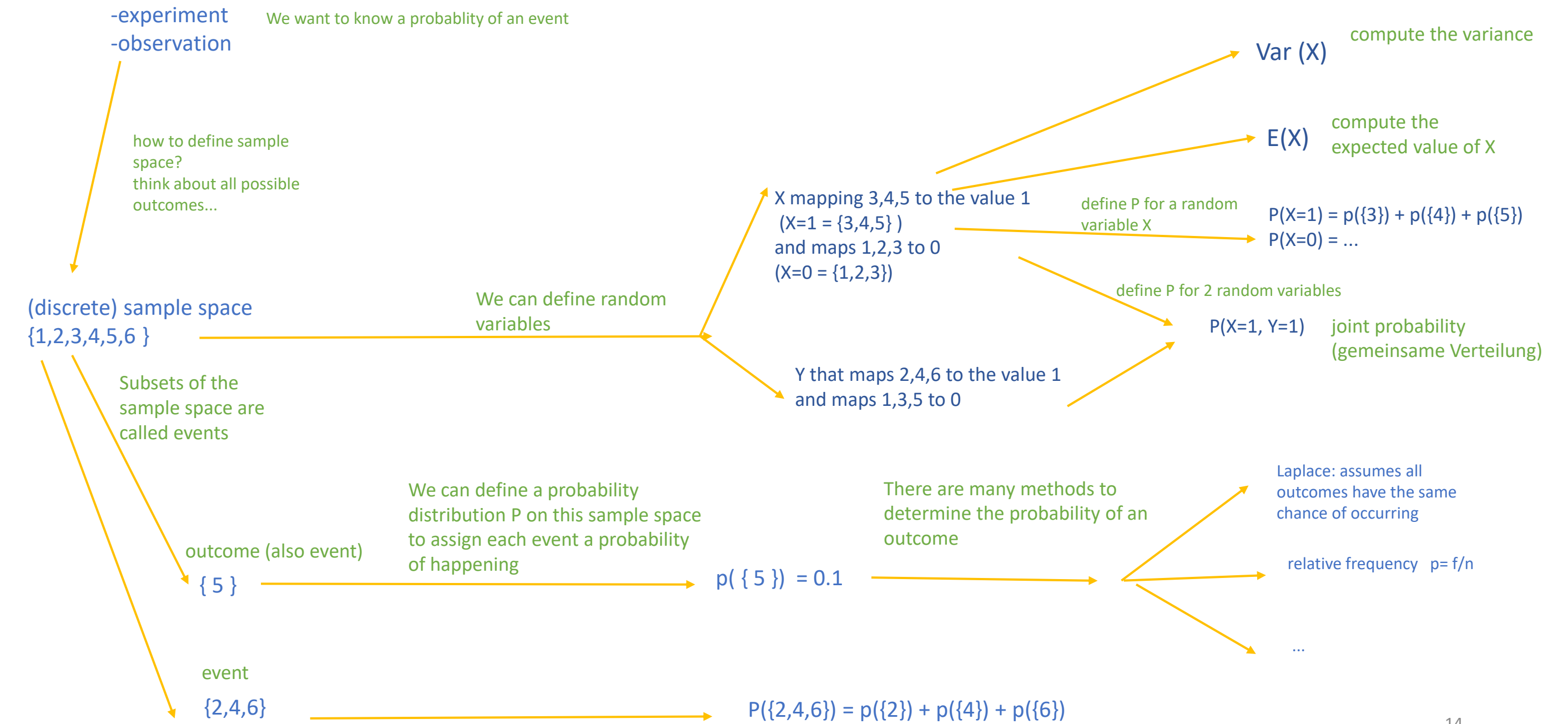
36 outcomes in total

Define a **random variable S** as „the sum of both dice“ (map the outcome to the sum of it), we get

	Possibilities
$P(S = 2)$	$= \frac{1}{36}$ (1, 1)
$P(S = 3)$	$= \frac{2}{36}$ (1, 2), (2, 1)
$P(S = 4)$	$= \frac{3}{36}$ (1, 3), (3, 1), (2, 2)
$P(S = 5)$	$= \frac{4}{36}$ (1, 4), (4, 1), (2, 3), (3, 2)
$P(S = 6)$	$= \frac{5}{36}$ (1, 5), (5, 1), (2, 4), (4, 2), (3, 3)
$P(S = 7)$	$= \frac{6}{36}$ (1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)
$P(S = 8)$	$= \frac{5}{36}$ (2, 6), (6, 2), (3, 5), (5, 3), (4, 4)
$P(S = 9)$	$= \frac{4}{36}$ (3, 6), (6, 3), (4, 5), (5, 4)
$P(S = 10)$	$= \frac{3}{36}$ (4, 6), (6, 4), (5, 5)
$P(S = 11)$	$= \frac{2}{36}$ (5, 6), (6, 5)
$P(S = 12)$	$= \frac{1}{36}$ (6, 6)



#summary



Aufgabe 2) Was ist eine **Zufallsvariable**? Was haben Zufallsvariablen mit der Berechnung eines Notendurchschnittes zu tun? (Denken Sie an die Noten sehr gut, gut, befriedigend, ausreichend, mangelhaft, ungenügend.) (2 Punkte)

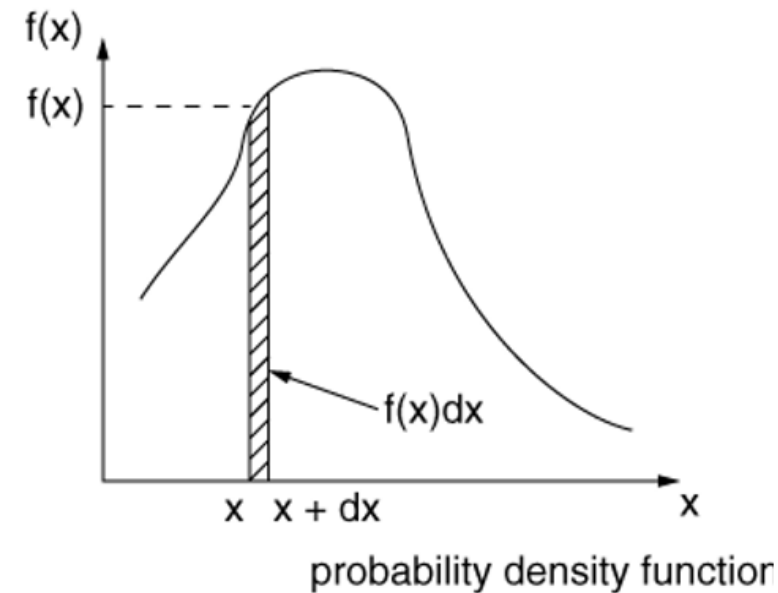
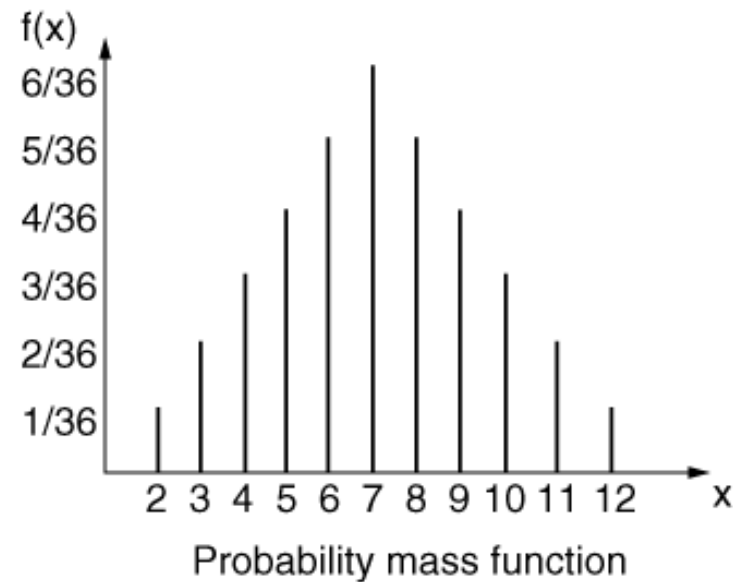
Aufgabe 2) Was ist eine **Zufallsvariable**? Was haben Zufallsvariablen mit der Berechnung eines Notendurchschnittes zu tun? (Denken Sie an die Noten sehr gut, gut, befriedigend, ausreichend, mangelhaft, ungenügend.) (2 Punkte)

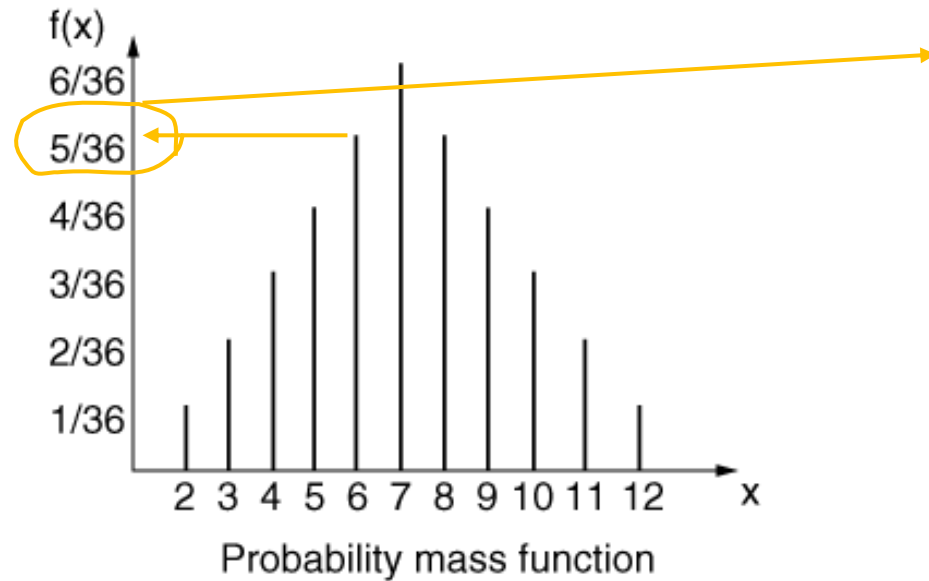
Zufallsvariable: Funktion, welche jedem Ergebnis eine reelle Zahl zuweist.

Beispiel: Abbildung der Noten *sehr gut, gut, befriedigend, ausreichend, mangelhaft, ungenügend* auf die Zahlen 1, 2, 3, 4, 5, 6.

Discrete and continuous sample space

- A sample space may either be discrete or continuous.
 - discrete: having countable infinite number of basic outcomes (e.g. counting the points on a die, value of a coin)
 - continuous: having an uncountable number of basic outcomes (e.g. measuring the height of a person, temperature)
- When a random variable is discrete, we use the **probability mass function (PMF)** to describe its probability distribution.
- When a random variable is continuous, we use the **probability density function (PDF)** to describe its probability distribution.



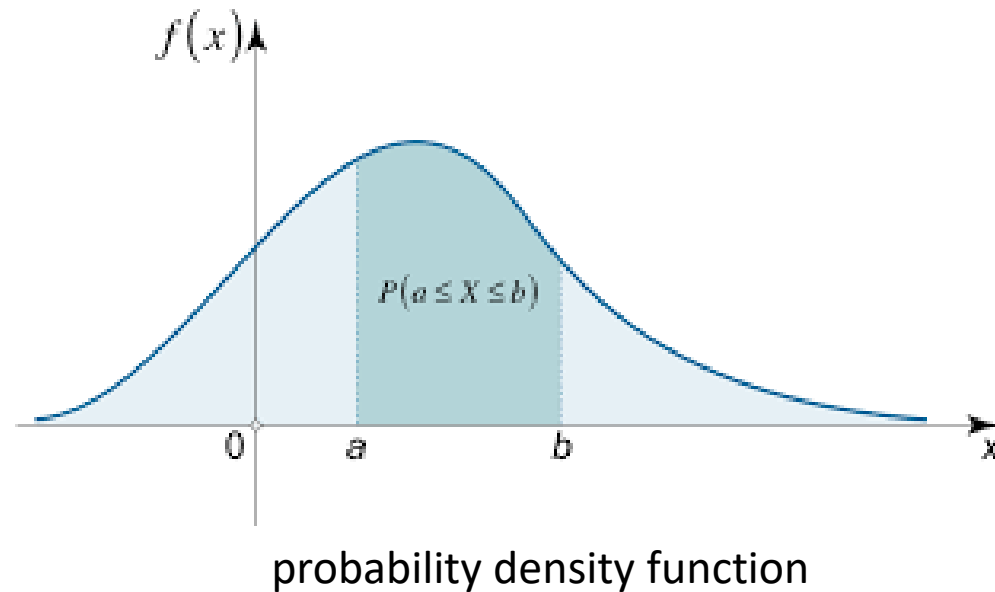


This is the probability of 6

- For the probability density function, we do not compute a probability of a specific value of a random variable, but for an interval.
- To get the probability of an interval a and b , we compute the area under the graph for this interval.
- You don't need to know how to compute this, because it is not important for our course. But if you want to know, check out this video

How to compute the area under the graph:

https://www.youtube.com/watch?v=hDjcx9p0ak&ab_channel=IntelligentSystemsLab



Formal definition of PMF and PDF (*not relevant for us)

Definition 3.1

Let X be a discrete random variable with range $R_X = \{x_1, x_2, x_3, \dots\}$ (finite or countably infinite). The function

$$P_X(x_k) = P(X = x_k), \text{ for } k = 1, 2, 3, \dots,$$

is called the *probability mass function (PMF)* of X .

Definition 4.2

Consider a continuous random variable X with an absolutely continuous CDF $F_X(x)$. The function $f_X(x)$ defined by

$$f_X(x) = \frac{dF_X(x)}{dx} = F'_X(x), \quad \text{if } F_X(x) \text{ is differentiable at } x$$

is called the *probability density function (PDF)* of X .

Bedingte Wahrscheinlichkeit

If we know that B has already occurred, what is the probability that A will also occur?

Bedingte Wahrscheinlichkeit: Wahrscheinlichkeit eines Ereignisses A, wenn das Ereignis B bekannt ist:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

\cap ist der Schnittmengenoperator.

the event we are interested in

Beispiel: Wahrscheinlichkeit $P(A)$, dass die Augenzahl eines Würfels gerade ist,
wenn die Augenzahl größer als 3 ist:

already happened/known

Bedingte Wahrscheinlichkeit

If we know that B has already occurred, what is the probability that A will also occur?

Bedingte Wahrscheinlichkeit: Wahrscheinlichkeit eines Ereignisses A, wenn das Ereignis B bekannt ist:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

\cap ist der Schnittmengenoperator.

the event we are interested in (A)

Beispiel: Wahrscheinlichkeit $P(A)$, dass die Augenzahl eines Würfels gerade ist,
wenn die Augenzahl größer als 3 ist:
already happened (B)

$$P(A) = \frac{p(4) + p(6)}{p(4) + p(5) + p(6)} = \frac{2/6}{3/6} = \frac{2}{3}$$

How to solve this question:

- What is A and B?
 - $A = \{2, 4, 6\}$
 - $B = \{3, 4, 5, 6\}$
- What is $P(B)$?
 - $P(B) = p(3) + p(4) + p(5) = 1/6 * 3 = 3/6$
- What is $P(A \text{ intersect } B)$?
 - $= P(\{4, 6\}) = 1/6 * 2 = 2/6$
- $P(A|B) = (1/6) / (3/6)$

Die **gemeinsame Verteilung** zweier Zufallsvariablen X und Y :

$$p(x, y) = p(X=x, Y=y) = P(A_x \cap A_y)$$

Example: In a die-rolling experiment, we define the following.

- a random variable X that maps all outcomes that are **greater than 3** to the value 1, and other outcomes to 0
- a random variable Y that maps all outcomes that are **even** to the value 1, and the rest to 0.
- What is the probability of getting the outcome that is greater than 3 and also even?

Solution

- We want to compute $P(X=1, Y=1)$
- see $X=1$ and $Y=1$ as sets and compute the intersection of them, then compute P
 - $X=1 : \{4,5,6\}$, $Y=1 : \{2,4,6\}$
 - $\{4,5,6\} \text{ intersect } \{2,4,6\} = \{4,6\}$
 - $P(\{4,6\}) = p(\{4\}) + p(\{6\}) = 1/6 * 2 = 2/6$

Beispiel:

AZ	X (AZ>3)	Y (AZ gerade)
1	0	0
2	0	1
3	0	0
4	1	1
5	1	0
6	1	1

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

Randverteilungen

Aus der gemeinsamen Verteilung der Zufallsvariablen X und Y , kann man die Verteilungen von X und Y berechnen.

Man nennt diese dann die **Randverteilungen**:

$$p_X(x) = \sum_{y \in \Omega_Y} p(x, y) \qquad p_Y(y) = \sum_{x \in \Omega_X} p(x, y)$$

Oft schreibt man statt $p_X(x)$ einfach $p(x)$.

Marginal Probability

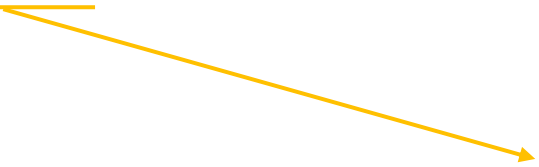
Example: In a die-rolling experiment, we define the following.

- a random variable X that maps all outcomes that are **greater than 3** to the value 1, and other outcomes to 0
- a random variable Y that maps all outcomes that are **even** to the value 1, and the rest to 0.
- What is the **marginal probability** $P_X(1)$?

Solution

- To compute $P_X(1)$, we have to know $P(X=1, Y=1)$ and $p(X=1, Y=0)$ so that we can compute $p(x,y)$ in the formula below.

$$p_X(x) = \sum_{y \in \Omega_Y} p(x, y)$$



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

- $P_X(1) = P(X=1, Y=1) + p(X=1, Y=0)$
= $2/6 + 1/6$
= $3/6$
= 0.5

Beispiel:	AZ	X	Y	X	Y	$p(x, y)$	$p_X(x)$	$p_Y(y)$
	1	0	0	0	0	2/6	0	3/6
	2	0	1	0	1	1/6	1	3/6
	3	0	0	1	0	1/6		
	4	1	1	1	1	2/6		
	5	1	0					
	6	1	1					

https://www.youtube.com/watch?v=SrEmzdOT65s&ab_channel=zedstatistics

Basic probability: Joint, marginal and conditional probability | Independence

	Male	Female	TOTAL
Game of thrones	80	120	200
West World	100	25	125
Other	50	125	175
TOTAL	230	270	500

	Male	Female	TOTAL
Game of thrones	0.16	0.24	0.4
West World	0.2	0.05	0.25
Other	0.1	0.25	0.35
TOTAL	0.46	0.54	1

Marginal probability distribution

Statistische Unabhängigkeit

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

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

Beispiel: Wurf zweier Würfel (analog für andere Würfelergebnisse)

$$p(W_1 = 1, W_2 = 4) = p(W_1 = 1) \cdot p(W_2 = 4) = 1/6 \cdot 1/6$$

Gegenbeispiel: Augenzahl > 3 ($=X$) und Augenzahl gerade ($=Y$)

$$p(X = 0, Y = 1) = 1/6 \neq 1/4 = p(X = 0) \cdot p(Y = 1)$$

We have to show that the left side is equal to the right side for each equation listed here.

$p(W_1=1, W_2=1) = ? p(W_1=1) * p(W_2=1)$
 $p(W_1=1, W_2=2) = ? p(W_1=1) * p(W_2=2)$
 $p(W_1=1, W_2=3) = ? p(W_1=1) * p(W_2=3)$
 $p(W_1=1, W_2=4) = ? p(W_1=1) * p(W_2=4)$
 $p(W_1=1, W_2=5) = ? p(W_1=1) * p(W_2=5)$
 $p(W_1=1, W_2=6) = ? p(W_1=1) * p(W_2=6)$
 $p(W_1=2, W_2=1) = ? p(W_1=2) * p(W_2=1)$
 $p(W_1=2, W_2=2) = ? p(W_1=2) * p(W_2=2)$
 $p(W_1=2, W_2=3) = ? p(W_1=2) * p(W_2=3)$
 $p(W_1=2, W_2=4) = ? p(W_1=2) * p(W_2=4)$
 $p(W_1=2, W_2=5) = ? p(W_1=2) * p(W_2=5)$
 $p(W_1=2, W_2=6) = ? p(W_1=2) * p(W_2=6)$
 $p(W_1=3, W_2=1) = ? p(W_1=3) * p(W_2=1)$
 $p(W_1=3, W_2=2) = ? p(W_1=3) * p(W_2=2)$
 $p(W_1=3, W_2=3) = ? p(W_1=3) * p(W_2=3)$
 $p(W_1=3, W_2=4) = ? p(W_1=3) * p(W_2=4)$
 $p(W_1=3, W_2=5) = ? p(W_1=3) * p(W_2=5)$
 $p(W_1=3, W_2=6) = ? p(W_1=3) * p(W_2=6)$
 $p(W_1=4, W_2=1) = ? p(W_1=4) * p(W_2=1)$
 $p(W_1=4, W_2=2) = ? p(W_1=4) * p(W_2=2)$
 $p(W_1=4, W_2=3) = ? p(W_1=4) * p(W_2=3)$
 $p(W_1=4, W_2=4) = ? p(W_1=4) * p(W_2=4)$
 $p(W_1=4, W_2=5) = ? p(W_1=4) * p(W_2=5)$
 $p(W_1=4, W_2=6) = ? p(W_1=4) * p(W_2=6)$
 $p(W_1=5, W_2=1) = ? p(W_1=5) * p(W_2=1)$
 $p(W_1=5, W_2=2) = ? p(W_1=5) * p(W_2=2)$
 $p(W_1=5, W_2=3) = ? p(W_1=5) * p(W_2=3)$
 $p(W_1=5, W_2=4) = ? p(W_1=5) * p(W_2=4)$
 $p(W_1=5, W_2=5) = ? p(W_1=5) * p(W_2=5)$
 $p(W_1=5, W_2=6) = ? p(W_1=5) * p(W_2=6)$
 $p(W_1=6, W_2=1) = ? p(W_1=6) * p(W_2=1)$
 $p(W_1=6, W_2=2) = ? p(W_1=6) * p(W_2=2)$
 $p(W_1=6, W_2=3) = ? p(W_1=6) * p(W_2=3)$
 $p(W_1=6, W_2=4) = ? p(W_1=6) * p(W_2=4)$
 $p(W_1=6, W_2=5) = ? p(W_1=6) * p(W_2=5)$
 $p(W_1=6, W_2=6) = ? p(W_1=6) * p(W_2=6)$

Erwartungswert

to compute this we have to know $p(x)$ for every x (value of X).

Der **Erwartungswert** ist der Mittelwert einer Zufallsvariablen:

$$E(X) = \sum_{x \in \Omega_X} p(x)x$$

Example1: compute the expected value of the following X

$X=1$ with $P(X=1)$ or $p(1) = 0.7$

$X=0$ with $P(X=0)$ or $p(0) = 0.3$

$$\begin{aligned} E(X) &= p(1)*1 + p(0) * 0 \\ &= 0.7*1 + 0.3 * 0 \\ &= 0.7 + 0 \\ &= 0.7 \end{aligned}$$

Example2: In a die rolling experiment with 6 outcomes, X maps each outcome to its digit value (e.g. Result 6 -> 6). Compute $E(X)$

- think about how many value X has and what is the probability of each value x .

sample space = $\{1,2,3,4,5,6\}$

$X=1 = \{1\}$, $p(1) = 1/6$

$X=2 = \{2\}$, $p(2) = 1/6$

..

$X=6 = \{6\}$, $p(6) = 1/6$

$$\begin{aligned} E(X) &= p(1)*1 + p(2) * 2 + p(3)*3 + p(4)*4 + p(5)*5 + p(6)*6 \\ &= 1/6 (1+2+3+4+5+6) \end{aligned}$$

<https://goodcalculators.com/expected-value-calculator/>

Uses and applications [\[edit \]](#)

The expectation of a random variable plays an important role in a variety of contexts. For example, in [decision theory](#), an agent making an optimal choice in the context of [incomplete information](#) is often assumed to maximize the expected value of their [utility function](#). For a different example, in [statistics](#), where one seeks estimates for unknown parameters based on available data, [the estimate itself is a random variable](#). In such settings, a desirable criterion for a "good" estimator is that it is [unbiased](#); that is, the expected value of the estimate is equal to the true value of the underlying parameter.

Expected values can also be used to compute the [variance](#), by means of the computational formula for the variance

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

A [very important application](#) of the expectation value is in the field of [quantum mechanics](#). The expectation value of a quantum mechanical operator \hat{A} operating on a [quantum state](#) vector $|\psi\rangle$ is written as $\langle \hat{A} \rangle = \langle \psi | \hat{A} | \psi \rangle$. The [uncertainty](#) in \hat{A} can be calculated using the formula $(\Delta A)^2 = \langle \hat{A}^2 \rangle - \langle \hat{A} \rangle^2$.

Erwartungswert einer Funktion f :

$$E(\underline{f(X)}) = \sum_{x \in \Omega_X} p(x) \underline{f(x)}$$

Example

Given X with $P(X=1) = 0.2$, $P(X=0) = 0.6$, $P(X=2) = 0.2$

What is $E(X^2)$?

Solution

- see X^2 as a function of x , meaning $f(x) = x^2$

- compute $f(x)$ for every x

$$f(1) = 1^2 = 1$$

$$f(0) = 0^2 = 0$$

$$f(2) = 2^2 = 4$$

- we already know $p(x)$ for every x , so we can now compute E

$$E(X^2) = p(1) * f(1) + p(0) * f(0) + p(2) * f(2)$$

$$= 0.2 * 1 + 0.6 * 0 + 0.2 * 4$$

Die **Varianz** ist ein Maß dafür, wie stark die einzelnen Werte vom Mittelwert abweichen:

$$= E[X^2] - E[X]^2$$

$$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E^2(X)$$

The **variance** is a measure of **variability**. It is calculated by taking the average of squared deviations from the mean.

Variance tells you the degree of spread in your data set. The more spread the data, the larger the variance is in relation to the **mean**.

High Variance

Metric Lift: 0.30
P-Value: 0.06689



vs.

Low Variance

Metric Lift: 0.30
P-Value: 0.04691



Die **Varianz** ist ein Maß dafür, wie stark die einzelnen Werte vom Mittelwert abweichen:

$$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E^2(X)$$

Example: Rolling a die with 6 outcomes and a uniform probability distribution. What is $\text{Var}(X)$?
(X maps each outcome to its numerical value e.g. one \rightarrow 1, two \rightarrow 2, ...)

we want to compute

$$E((X - E(X))^2) =$$

See this as a function of x

$$f(x) = [x - E(X)]^2$$

We know that $E(X) = 3.5$

So we have, $f(x) = [x - 3.5]^2$

- What are the values of X ?

$$X=1, f(1) = [1 - 3.5]^2$$

$$X=2, f(2) = [2 - 3.5]^2$$

$$X=3, f(3) = [3 - 3.5]^2$$

$$X=4, f(4) = [4 - 3.5]^2$$

$$X=5, f(5) = [5 - 3.5]^2$$

$$X=6, f(6) = [6 - 3.5]^2$$

All $p(x) = 1/6$

Erwartungswert einer Funktion f :

$$E(f(X)) = \sum_{x \in \Omega_X} p(x)f(x)$$

We know $p(x)$ and $f(x)$ for every x , so we can now compute $E(f(x))$

$$E(f(X)) = p(1)*f(1) + p(2)*f(2) + p(3)*f(3) + p(4)*f(4) + p(5)*f(5) + p(6)*f(6)$$

$$= 1/6 * [1 - 3.5]^2 + 1/6 * [2 - 3.5]^2 + \dots$$

Check the answer of the previous slide: <https://www.rapidtables.com/calc/math/variance-calculator.html>

Probability	Data number
0.67	1
0.67	2
0.67	3
0.67	4
0.67	5
0.67	6

Variance: 2.91666667

Mean: 3.5

Standard deviation: 1.70782513

Calculation:

$$\mu = (0.67 \times 1 + 0.67 \times 2 + 0.67 \times 3 + 0.67 \times 4 + 0.67 \times 5 + 0.67 \times 6) / (0.67 + 0.67 + 0.67 + 0.67 + 0.67 + 0.67) = 3.5$$
$$\sigma^2 = (0.67 \times (1 - 3.5)^2 + 0.67 \times (2 - 3.5)^2 + 0.67 \times (3 - 3.5)^2 + 0.67 \times (4 - 3.5)^2 + 0.67 \times (5 - 3.5)^2 + 0.67 \times (6 - 3.5)^2) / (0.67 + 0.67 + 0.67 + 0.67 + 0.67 + 0.67) = 2.91666667$$

Die **Varianz** ist ein Maß dafür, wie stark die einzelnen Werte vom Mittelwert abweichen:

$$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E^2(X)$$

Example: Given X with $P(X=3) = 0.7$ and $P(X=2) = 0.3$, what is the variance of X ?

We want to compute the variance $E((X - E(X))^2)$

$$E(f(X)) = \sum_{x \in \Omega_X} p(x)f(x)$$

- to compute the variance we need to know $E(X)$, $p(x)$, and $f(x)$ which is $[x - E(x)]^2$.
- first compute $E(X)$. For this, we need to know $p(x)$ and x
 - $E(X) = 3 \cdot 0.7 + 2 \cdot 0.3$
 - $= 2.1 + 0.6 = 2.7$
- compute $f(x)$
 - $X=3$, $f(3) = [3 - 2.7]^2$
 - $X=2$, $f(2) = [2 - 2.7]^2$

$$E(X) = \sum_{x \in \Omega_X} p(x)x$$

$$\begin{aligned} E((X - E(X))^2) &= p(3)f(3) + p(2)f(2) \\ &= 0.7 \cdot [3 - 2.7]^2 + 0.3 \cdot [2 - 2.7]^2 \end{aligned}$$

Check the answer of the previous slide: <https://www.rapidtables.com/calc/math/variance-calculator.html>

Discrete random variable variance calculator

Enter probability or weight and data number in each row:

Probability	Data number
0.7	3
0.3	2

Variance: 0.21

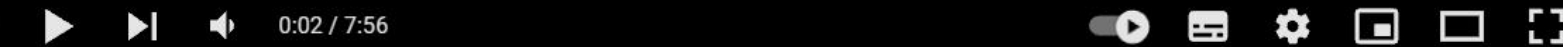
Mean: 2.7

Standard deviation: 0.45825757

Calculation:

$\mu = (0.7 \times 3 + 0.3 \times 2) / (0.7 + 0.3) = 2.7$
 $\sigma^2 = (0.7 \times (3 - 2.7)^2 + 0.3 \times (2 - 2.7)^2) / (0.7 + 0.3) = 0.21$

Expectation and Variance of Discrete Random Variables



Expected Value and Variance of Discrete Random Variables

https://www.youtube.com/watch?v=OvTEhNL96v0&ab_channel=jbstatics

Die **Standardabweichung** ist die Wurzel aus der Varianz.

$$SD = \sqrt{Var(X)}$$

Variance vs standard deviation

The **standard deviation** is derived from variance and tells you, on average, how far each value lies from the mean. It's the square root of variance.

Both measures reflect **variability** in a distribution, but their units differ:

- **Standard deviation** is expressed in the same units as the original values (e.g., meters).
- **Variance** is expressed in much larger units (e.g., meters squared)

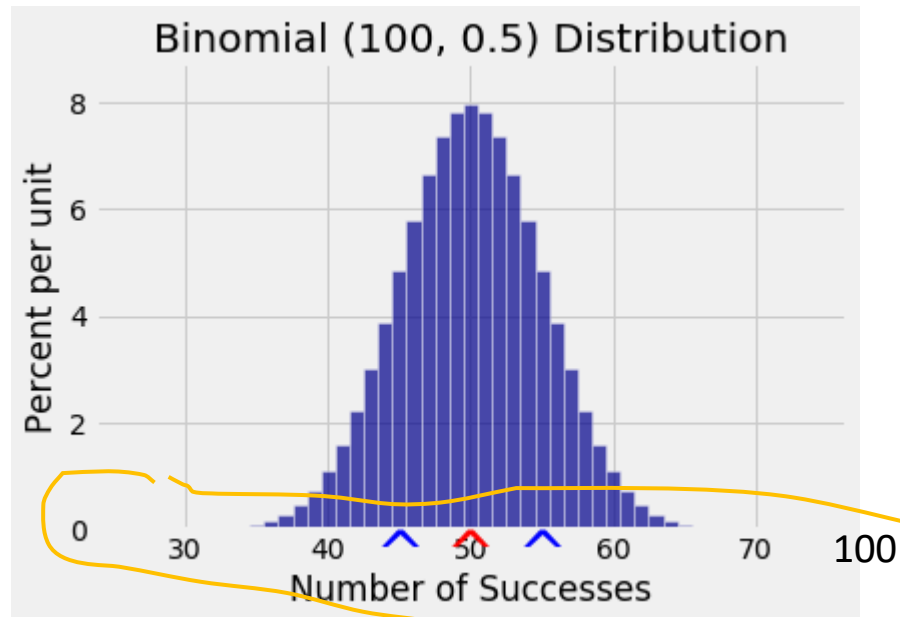
Since the units of variance are much larger than those of a typical value of a data set, it's harder to interpret the variance number intuitively. That's why standard deviation is often preferred as a main measure of variability.

However, the variance is more informative about variability than the standard deviation, and it's used in making **statistical inferences**.

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 (heads, tails) 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 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.

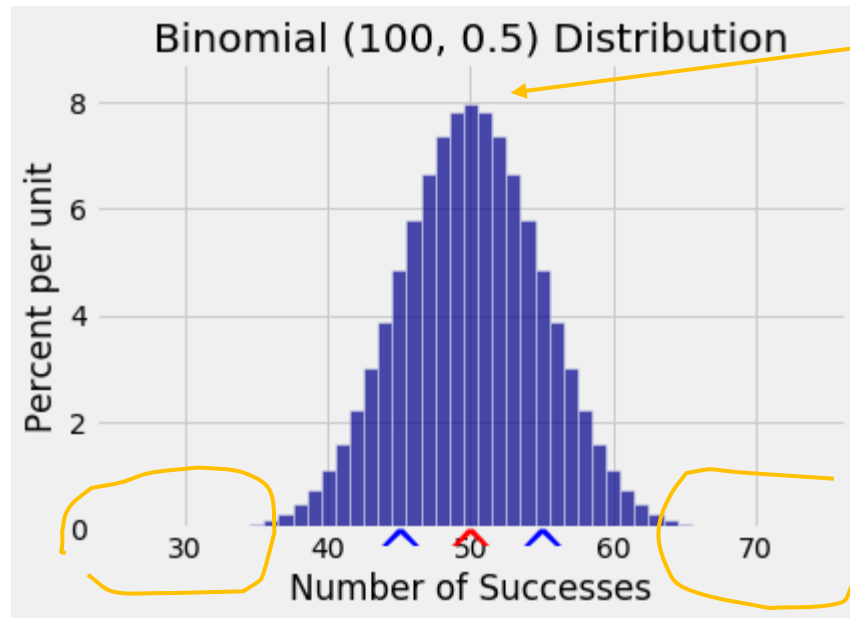
This is the value of a new random variable for which we define the binomial distribution. Let's call it B .

$P(B=0)$ 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.

...

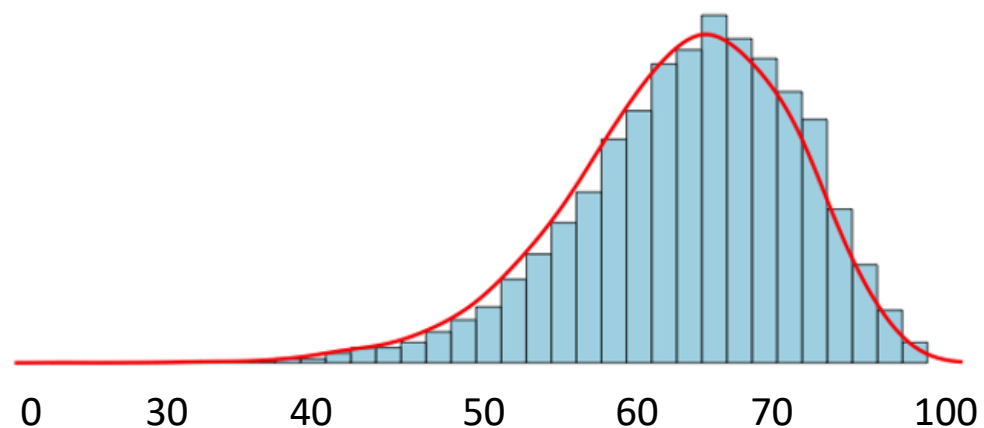
$P(B=100)$ is the probability that we get head 100 time out of 100.



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 $P(\text{head})$ is not 0.5, but higher, for example, $p(\text{head}) = 0.7$
The binomial graph might look different.



Formula for calculating a binomial probability

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

We need r , n , p to compute this.

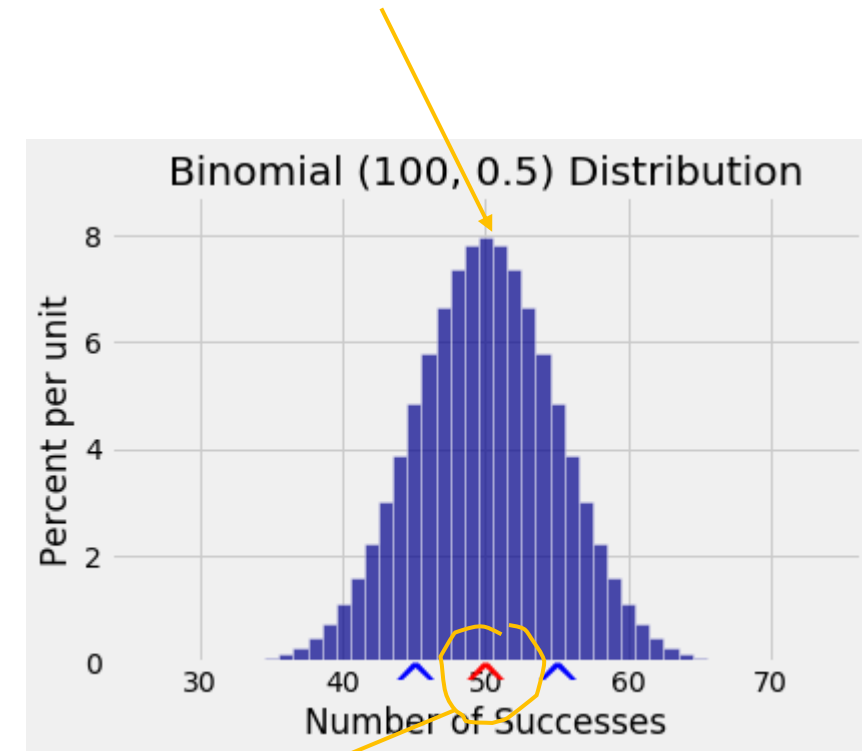
r is the number of time that head occurs

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.

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

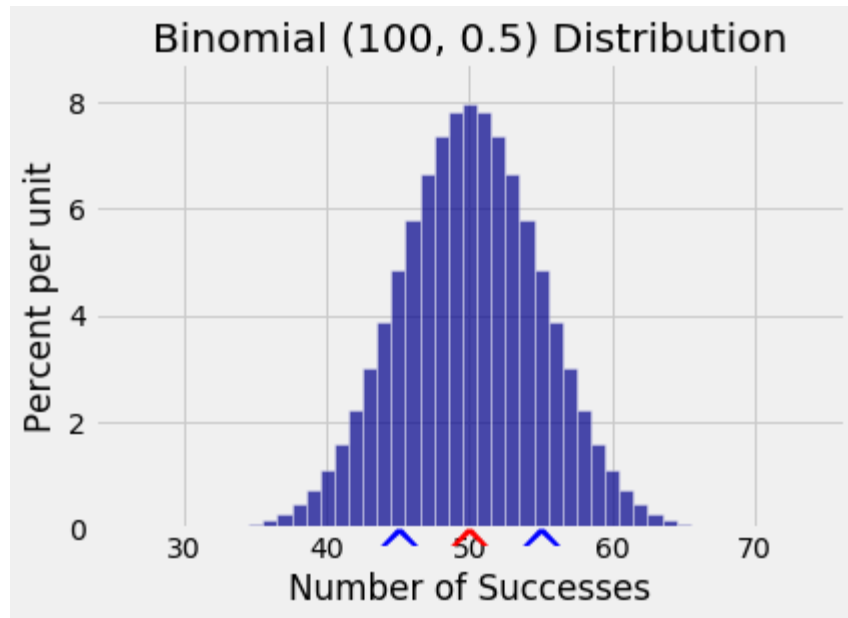


r

Statistical hypothesis testing

Situation: You know that if the coin is fair, $p(\text{heads})$ must be 0.5, and if you plot a binomial distribution, you 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.



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 $P(\text{heads})=0.5$ and $P(\text{tails})=0.5$. Maybe $P(\text{heads}) > 0.5$.

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 a decision about p .

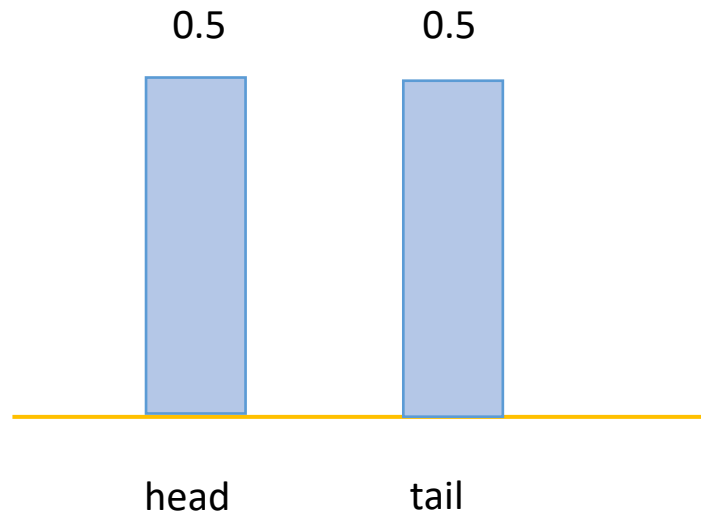
And even though the probability of getting head 80 times is very low in this situation, it is not zero. This means that there is still a small chance to get such a result even if $p(\text{head})=0.5$.

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**.

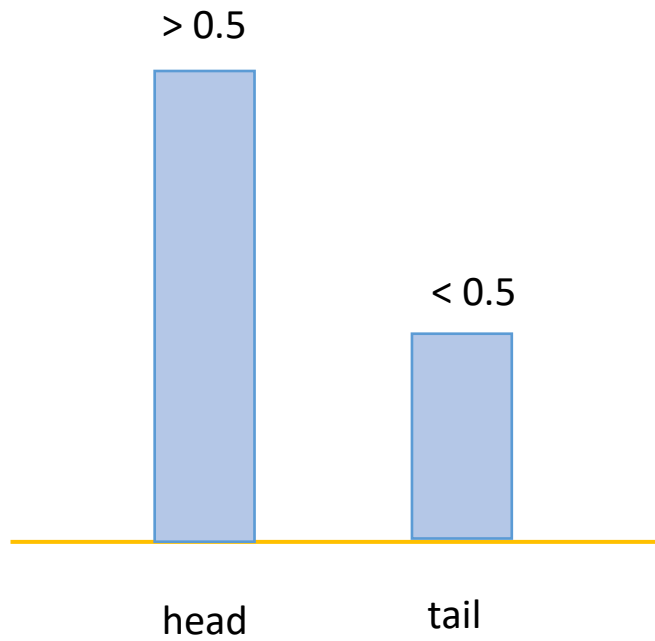
the binomial test

H₀ : null hypothesis saying $p(\text{head}) = 0.5$

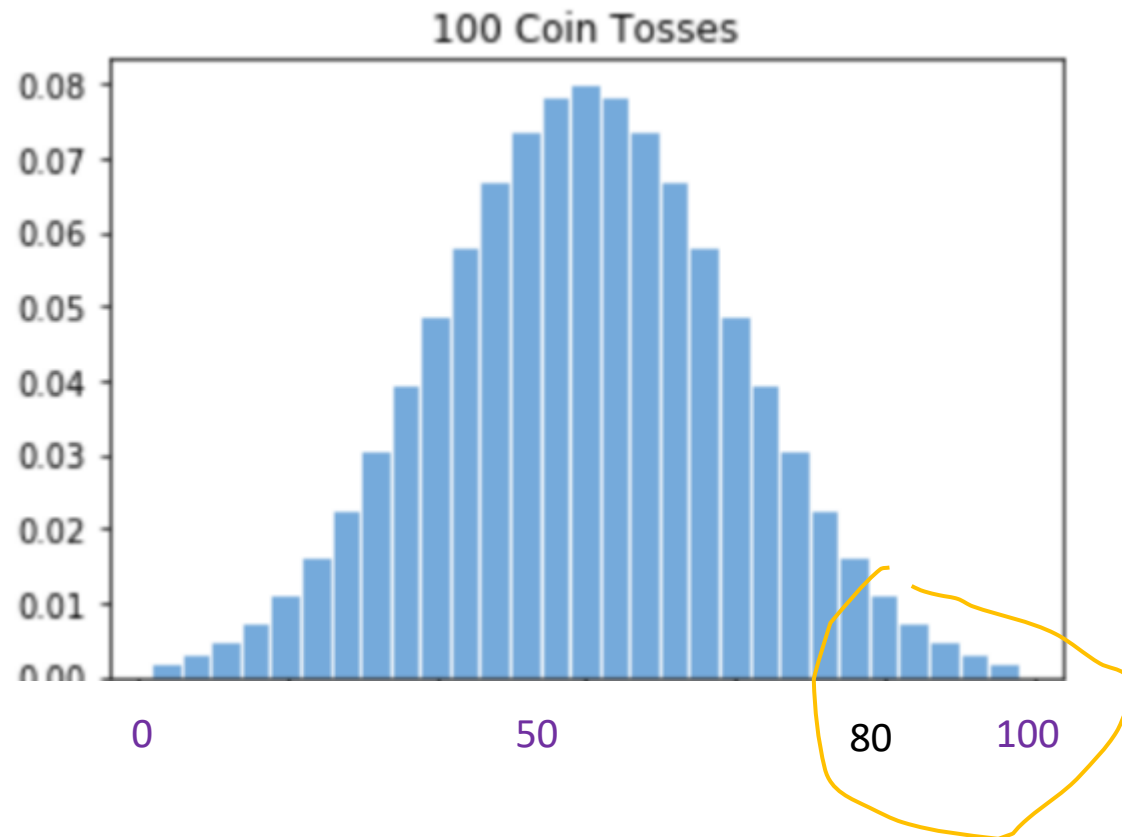


You suspect that $p(\text{head})$ might be greater than 0.5

H_{alternative}: $p(\text{head}) > 0.5$



the binomial test



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 we first assumed. In other words, is it low enough to say that $p(\text{head})$ is not 0.5 but more? This probability is written as $p(B \geq 80)$.

You can think of it this way, if $p(B \geq 80)$ is super low, it means getting head 80 times does not seem like a normal thing to occur under the null hypothesis which says $p(\text{head})=0.5$.

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

Now you have to pick a significant value. If you pick 0.05 it means you will only believe your alternative hypothesis (which says $p(\text{head}) > 0.5$) only if $p(B \geq 80)$ 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 the your test result)

For example, if $p(B \geq 80) = 0.0123$ which is less than 0.05. You can reject the null hypothesis.

If $p(B \geq 80) = 0.0567$, you can not reject the null hypothesis, because the probability is greater than 0.05.

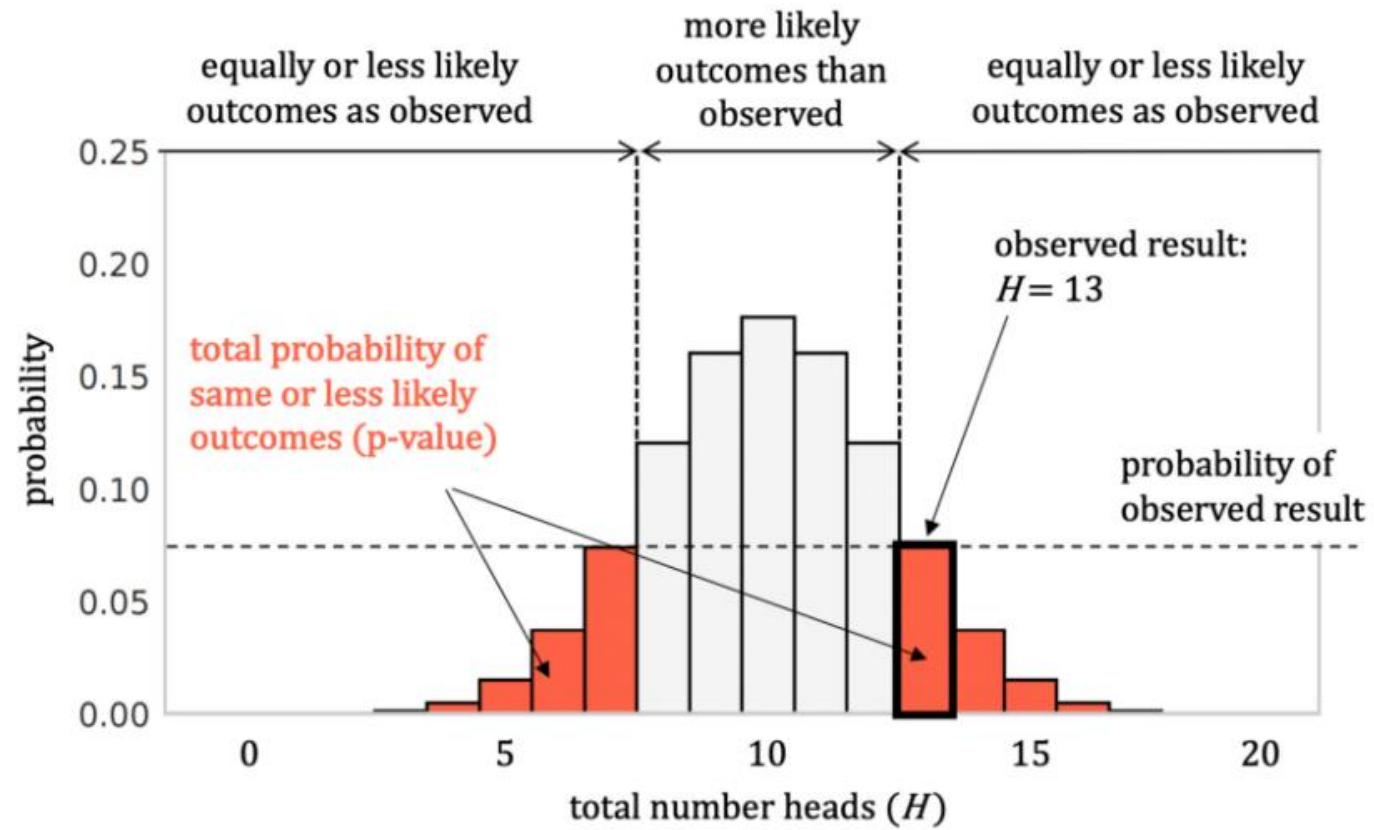


Figure 1. Probability distribution for all possible outcomes of 20 coin flips assuming the coin is fair and $P(\text{heads}) = 0.5$. Function is represented by Eq. 1 with parameters $\text{Bin}(H, N_{\text{tosses}} = 20, p = 0.5)$. The area in red corresponds to the total probability of observing an 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 bias? and bias toward head or tail? Prove this using a binomial test.
2. 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 a binomial test.

Solution:

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

$H_0 : P(\text{head}) = 0.5$

$H_{\text{Alternative}}: P(\text{head}) < 0.5$ (we suspect that the coin is biased toward tail)

compute $P_{\text{binomial}}(X \leq 10)$ which is $P_{\text{binomial}}(X=0) + P_{\text{binomial}}(X=1) + P_{\text{binomial}}(X=2) + \dots + P_{\text{binomial}}(X=10)$

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

If $P_{\text{binomial}}(X \leq 10) < 0.05$, then we reject H_{null} (meaning the test suggests that the coin is biased toward tail), otherwise we accept H_{null} (meaning the test suggests that the coin is not biased).

note: In the exam, you don't have to actually compute **$P_{\text{binomial}}(X \leq 10)$** because it would take a long time. 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

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,

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

Binomial Probability Calculator

Trials (n): *

Probability (p): *

Successes (X): *

Type of probability: *

At most X successes

CALCULATE

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 a binomial test.

Pick significant value = 0.05

$H_0 : P(\text{head}) = 0.5$

$H_{\text{Alternative}}: P(\text{head}) > 0.5$

compute $P_{\text{binomial}}(X \geq 8)$

$$P_{\text{binomial}}(X \geq 8) = P_{\text{binomial}}(X=8) + P_{\text{binomial}}(X=9) + P_{\text{binomial}}(X=10)$$

If $P_{\text{binomial}}(\geq 8) < 0.05$, 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).

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

Klausuraufgaben zu Binomialtest

Die folgenden Klausuren sind nur dazu da, die Anwendung des Binomialtests in NLP zu zeigen. Es wird nicht erwartet, dass Sie sie sofort lösen können. Sie sollten zunächst etwas über Tagger und Klassifikation lernen.

Aufgabe 3) Sie sollen untersuchen, ob TaggerA **signifikant** besser ist als der TaggerB. Auf Testdaten haben Sie folgende Ergebnisse erhalten:

Satz:	Ich	sah	den	Mann	auf	dem	Hügel	mit	dem	Stock	unter	dem	Arm
Tags:	PPER	VVFIN	ART	NN	APPR	ART	NN	APPR	ART	NN	APPR	ART	NN
TaggerA:	PPER	VVIN	ART	NNS	APPR	ART	NE	APPR	PDS	NN	APPO	PDS	NN
TaggerB:	NN	VVFIN	ART	NE	APPO	PDS	NN	APPR	PDS	NE	APPR	ART	NE

Führen Sie einen **Vorzeichen-Test** durch. Geben Sie dabei alle Zwischenschritte an.

Wie lautet die Nullhypothese?

(Werte der Binomialfunktion müssen Sie hier nicht ausrechnen.)

(5 Punkte)

Aufgabe 3) Sie sollen untersuchen, ob TaggerA **signifikant** besser ist als der TaggerB.

Auf Testdaten haben Sie folgende Ergebnisse erhalten:

Satz:	Ich	sah	den	Mann	auf	dem	Hügel	mit	dem	Stock	unter	dem	Arm
Tags:	PPER	VVFIN	ART	NN	APPR	ART	NN	APPR	ART	NN	APPR	ART	NN
TaggerA:	PPER	VVINP	ART	NNS	APPR	ART	NE	APPR	PDS	NN	APPO	PDS	NN
TaggerB:	NN	VVFIN	ART	NE	APPO	PDS	NN	APPR	PDS	NE	APPR	ART	NE

Führen Sie einen **Vorzeichen-Test** durch. Geben Sie dabei alle Zwischenschritte an.

Wie lautet die Nullhypothese?

(Werte der Binomialfunktion müssen Sie hier nicht ausrechnen.)

(5 Punkte)

Lösung

Wir müssen zählen, wieviele Wörter nur TaggerA korrekt annotiert hat, und wieviele Wörter nur TaggerB richtig annotiert hat.

nur TaggerA korrekt: 5

nur TaggerB korrekt: 4

Wir haben also 9 Beispiele, die genau ein Tagger korrekt annotiert hat. Wir nehmen (nicht ganz korrekt) an, dass diese Beispiele eine Stichprobe von statistisch unabhängigen Ergebnissen bildet.

Nullhypothese: TaggerA ist nicht besser als TaggerB

Bei jedem Element der Stichprobe ist die Wahrscheinlichkeit, dass TaggerA richtig lag, maximal 0.5.

Wir summieren die Werte der Binomialfunktion für r-Werte ab 5: $p = \sum_{r=5}^9 b(r, 0.5, 9)$

TaggerA ist signifikant besser als TaggerB, falls $p \leq 0.05$ gilt.

Aufgabe 9) Angenommen Sie vergleichen Ihren neu entwickelten Spam-Klassifizierer *Spammy* mit einem Baseline-Spam-Klassifizierer und erhalten folgende Ergebnisse:

Goldstandard	Baseline	Spammy	Häufigkeit
Spam	Spam	Spam	57
Spam	Spam	NoSpam	5
Spam	NoSpam	Spam	7
Spam	NoSpam	NoSpam	3
NoSpam	Spam	Spam	2
NoSpam	Spam	NoSpam	6
NoSpam	NoSpam	Spam	2
NoSpam	NoSpam	NoSpam	154

Hier gab es bspw. 57 Emails, die sowohl vom Baseline-Tagger als auch von Spammy korrekt als Spam klassifiziert wurden.

Sagen Sie so genau wie möglich, wie Sie hier mit dem **Vorzeichentest** berechnen, ob Spammy signifikant besser als der Baseline-Klassifikator ist. (3 Punkte)

Antwort:

Man zählt zunächst, wieviele Emails Spammy richtig klassifiziert und der andere Tagger falsch. Das sind 13. Dann zählt man, wieviele Emails Spammy falsch klassifiziert und der andere Tagger richtig. Das sind 7. Nur diese 20 Beispiele sind für den Vorzeichentest relevant. Die Nullhypothese besagt, dass Spammy nicht besser als der andere Tagger ist. Die Wahrscheinlichkeit, dass Spammy ein beliebiges der 20 Beispiele korrekt klassifiziert hat, ist daher unter Annahme der Nullhypothese maximal 0.5. Die Wahrscheinlichkeit, dass man bei Gültigkeit der Nullhypothese das beobachtete Ergebnis (Spammy 13 Mal korrekt) oder ein noch unwahrscheinlicheres Ergebnis (≥ 13) bekommt, ist durch die Summe $\sum_{i=13}^{20} b(i, 0.5, 20)$ gegeben, wobei $b(r, p, n)$ die Binomialverteilung mit Wahrscheinlichkeit p und Stichprobengröße n ist. Wenn diese Summe kleiner als 0.05 ist, kann die Nullhypothese zurückgewiesen werden. Man sagt dann: Spammy hat eine signifikant höhere Genauigkeit.

Aufgabe 7) Erklären Sie, wie der Vorzeichentest (Binomialtest) funktioniert, den man zur Berechnung der Signifikanz beim Vergleich zweier Wortart-Tagger benutzen kann.
(3 Punkte)