Statistical methods in NLP
Probabilities and random variables

UNIVERSITY OF GOTHENBURG

Richard Johansson

January 22, 2016

today

- recap of a few probability notions, and two new ones
- random variables and their distributions

overview

recap: basic probability rules

two more basic probability rules

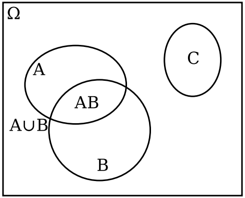
random variables and their distributions

mean and variance for random variables

the Bernoulli and binomial distributions

the mathematical definition: the Kolmogorov axioms

- the probability $P(A)$ is a number such that
  - $0 \leq P(A) \leq 1$ for every event A
  - $P(\Omega) = 1$
  - $P(A \cup B) = P(A) + P(B)$ if A and B are disjoint
- in the illustrations, $P(A)$ intuitively corresponds to the area covered by A in the Venn diagram

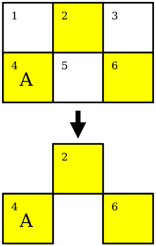


joint and conditional probabilities

- the probability of both A and B happening is called the **joint probability**, written $P(AB)$ or $P(A, B)$
- definition: if $P(B) \neq 0$, then

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

  is referred to as the **conditional probability of A given B**
- intuitively in the Venn diagram: zoom in on B
  - "what is the probability of a 4 if we know it's an even number?"



independent events

- definition: two events A and B are **independent** if

$$P(AB) = P(A) \cdot P(B)$$

- this can be rewritten in a more intuitive way: "the probability of A does not depend on anything about B"

$$P(A|B) = P(A)$$

overview

recap: basic probability rules

two more basic probability rules

random variables and their distributions

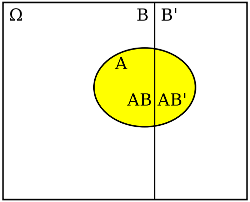
mean and variance for random variables

the Bernoulli and binomial distributions

the law of total probability

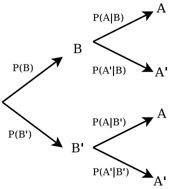
- from the definition of conditional probability, we get

$$P(AB) = P(A|B)P(B)$$



- we can do the same thing with B'

$$P(A\ B') = P(A|B')P(B')$$



- then

$$P(A) = P(AB) + P(A\ B')$$

$$= P(A|B)P(B) + P(A|B')P(B')$$

- this is a special case of the **law of total probability**

another example

$$P(\text{going bald}|\text{male}) = 0.4$$

$$P(\text{going bald}|\text{female}) = 0.01$$

$$P(\text{male}) = 0.49$$

$$P(\text{female}) = 0.51$$

$$P(\text{going bald}) =$$

another example

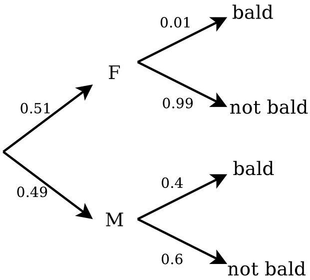
$$P(\text{going bald}|\text{male}) = 0.4$$

$$P(\text{going bald}|\text{female}) = 0.01$$

$$P(\text{male}) = 0.49$$

$$P(\text{female}) = 0.51$$

$$P(\text{going bald}) =$$



$$= P(\text{going bald}|\text{male}) \cdot P(\text{male}) + P(\text{going bald}|\text{female}) \cdot P(\text{female})$$

$$= 0.01 \cdot 0.49 + 0.4 \cdot 0.51 = 0.2089$$

Bayes' theorem

- in the NLP course, we already saw **Bayes' theorem**:

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

- this is often used to split a model into simpler parts

typical use of the Bayes theorem in NLP

- Bayes' theorem is involved in many NLP models
- the typical use is something like this (in this case, HMM tagging):

$$P(T|W) = \frac{P(W|T) \cdot P(T)}{P(W)}$$

- this trick is used in Naive Bayes classifiers, tagging, speech recognition, machine translation, and other applications
- often, the next step is the observation that we can simplify this if we're only interested in the maximum:

$$\begin{aligned}\arg\max_T P(T|W) &= \arg\max_T \frac{P(W|T) \cdot P(T)}{P(W)} \\ &= \arg\max_T P(W|T) \cdot P(T)\end{aligned}$$

how to get Bayes' theorem

- recall the definition of conditional probability

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

- we rearrange:

$$P(AB) = P(A|B) \cdot P(B)$$

- and by switching symbols:

$$P(AB) = P(B|A) \cdot P(A)$$

- by combining, we get Bayes' theorem

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

exercise: drug testing

- a drug test has a true positive rate of 99% and a true negative rate of 99%

$$P(\text{positive}|\text{user}) = 0.99 \quad P(\text{negative}|\text{not user}) = 0.99$$

- 0.5% of all people are users of the drug

$$P(\text{user}) = 0.005$$

- if a person tests positive, what is the probability that this is a user of a drug?

$$P(\text{user}|\text{positive}) = ?$$

exercise: drug testing (continued)

- idea: we use the given information and apply Bayes' theorem

exercise: drug testing (continued)

- idea: we use the given information and apply Bayes' theorem
- the missing piece for applying Bayes is $P(positive)$

exercise: drug testing (continued)

- idea: we use the given information and apply Bayes' theorem
- the missing piece for applying Bayes is $P(\textit{positive})$

$$
\begin{aligned}
P(\text{positive}) &= P(\text{positive}|\text{user}) \cdot P(\text{user}) \\
&\quad + P(\text{positive}|\text{not user}) \cdot P(\text{not user}) \\
&= 0.99 \cdot 0.005 + 0.01 \cdot 0.995 = 0.0149
\end{aligned}
$$

exercise: drug testing (continued)

- idea: we use the given information and apply Bayes' theorem
- the missing piece for applying Bayes is $P(\textit{positive})$

$$
\begin{aligned}
P(\text{positive}) &= P(\text{positive}|\text{user}) \cdot P(\text{user}) \\
&\quad + P(\text{positive}|\text{not user}) \cdot P(\text{not user}) \\
&= 0.99 \cdot 0.005 + 0.01 \cdot 0.995 = 0.0149
\end{aligned}
$$

- so finally:

$$
\begin{aligned}
P(\text{user}|\text{positive}) &= \frac{P(\text{user}|\text{positive}) P(\text{user})}{P(\text{positive})} \\
&= \frac{0.99 \cdot 0.005}{0.0149} = 0.332
\end{aligned}
$$

overview

recap: basic probability rules

two more basic probability rules

random variables and their distributions

mean and variance for random variables

the Bernoulli and binomial distributions

recap: random number generators in Python

- the two random number generating functions are examples of **random variables** with **uniform distributions**
  - this means that all outcomes are equally probable
  - if we generate a lot of random numbers, the histogram will be flat
- `random.randint(1, 6)` is a **discrete** uniform random variable
  - it generates 1, 2, 3, 4, 5, or 6 with equal probability $\frac{1}{6}$
- `random.random()` is a **continuous** uniform random variable
  - it generates any float between 0 and 1 with equal probability
- now: discrete random variables

random variables

- a **random variable** (r.v.) is a variable that selects its value randomly, like `random.randint` and `random.random`
  - also: **stochastic variable** (στοχαστικός)
- `random.randint` and `random.random` are uniform, but in general the different outcomes can have different probabilities
- examples:
  - the amount I win when buying a lottery ticket
  - the number of heads when tossing coins n times
  - the gender of a newborn baby
  - the number of words in an English sentence randomly selected from a corpus
  - the initial word in a random sentence

example: lottery

- my r.v. X is the amount of money I win when I buy a lottery ticket
- the possible outcomes:
  - if I win, I get 1,000,000 SEK
  - otherwise, I get nothing
- the probabilities of the outcomes:
  - $P(0 \text{ SEK}) = 0.99999$
  - $P(1,000,000 \text{ SEK}) = 0.00001$
  - $P(\text{something else}) = 0$

example: tossing a coin twice

- my r.v. X is the number of heads I get when tossing a coin twice
- the possible ways the coins can land:

  Head-Head, Head-Tail, Tail-Head, Tail-Tail

- assuming the coin is even, each of these possibilities has a probability of $\frac{1}{4}$
- so here are the probabilities for the different values of X:

$$P(X = 0) = \frac{1}{4}$$

$$P(X = 1) = \frac{2}{4}$$

$$P(X = 2) = \frac{1}{4}$$

describing a random variable

- when discussing a random variable, we need to describe which values it takes and with which probabilities: the **distribution**
- for instance:
  - when rolling a die, all the outcomes have the same probability

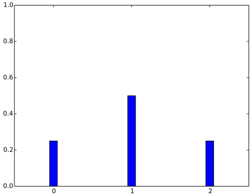
the probability mass function

- to describe the distribution of the r.v. X, we use a function called the **probability mass function** (pmf) of X:

$$p_X(x) = P(X \text{ takes the value } x)$$

- for instance, the number of heads when tossing a coin twice:

$$p_X(0) = P(X = 0) = \frac{1}{4}$$
$$p_X(1) = P(X = 1) = \frac{2}{4}$$
$$p_X(2) = P(X = 2) = \frac{1}{4}$$



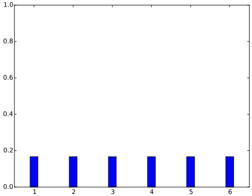
the pmf for a die roll

- the uniform distribution has a constant pmf:

$$p_X(1) = \frac{1}{6}$$

$$\dots$$

$$p_X(6) = \frac{1}{6}$$



how many times do I have to take the exam?

- the probability of passing the exam is 0.6
- if I fail, I don't prepare for the next one
- X = the number of times I have to take the exam to pass

how many times do I have to take the exam?

- the probability of passing the exam is 0.6
- if I fail, I don't prepare for the next one
- X = the number of times I have to take the exam to pass

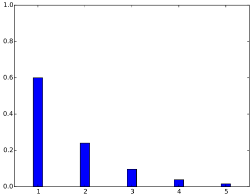
$$p_X(1) = 0.6$$

$$p_X(2) = 0.4 \cdot 0.6$$

$$p_X(3) = 0.4 \cdot 0.4 \cdot 0.6$$

$$\cdots$$

$$p_X(k) = 0.4^{(k-1)} \cdot 0.6$$



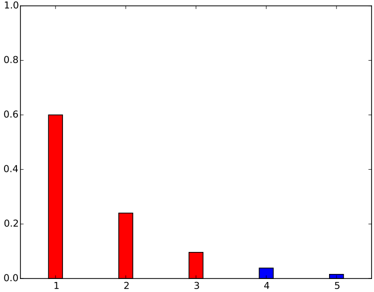
probabilities of intervals

- what is the probability that we'll go to the exam at most 3 times?

probabilities of intervals

- what is the probability that we'll go to the exam at most 3 times?

$$p_X(1) + p_X(2) + p_X(3) = 0.6 + 0.4 \cdot 0.6 + 0.4^2 \cdot 0.6$$



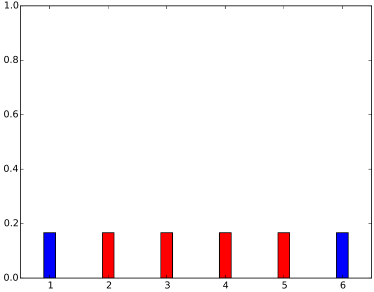
probabilities of intervals (2)

- what is the probability that we roll a number between 2 and 5?

probabilities of intervals (2)

- what is the probability that we roll a number between 2 and 5?

$$p_X(2) + p_X(3) + p_X(4) + p_X(5) = 4 \cdot \frac{1}{6}$$



overview

recap: basic probability rules

two more basic probability rules

random variables and their distributions

mean and variance for random variables

the Bernoulli and binomial distributions

recap: the mean of a sample

- recall that the sample mean $\bar{x}$ of a dataset x is defined

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- mean of `[2, 6, 1, 1, 5, 4, 6, 4, 1, 3]`:

recap: the mean of a sample

- recall that the sample mean $\bar{x}$ of a dataset x is defined

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- mean of `[2, 6, 1, 1, 5, 4, 6, 4, 1, 3]`:

$$\tfrac{1}{10}(2+6+1+1+5+4+6+4+1+3)$$

recap: the mean of a sample

- recall that the sample mean $\bar{x}$ of a dataset x is defined

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- mean of `[2, 6, 1, 1, 5, 4, 6, 4, 1, 3]`:

$$\frac{1}{10}(2+6+1+1+5+4+6+4+1+3)$$
$$= \frac{1}{10}(3 \cdot 1 + 1 \cdot 2 + 1 \cdot 3 + 2 \cdot 4 + 1 \cdot 5 + 2 \cdot 6)$$

recap: the mean of a sample

- recall that the sample mean $\bar{x}$ of a dataset x is defined

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- mean of `[2, 6, 1, 1, 5, 4, 6, 4, 1, 3]`:

$$\begin{aligned}
&\tfrac{1}{10}(2+6+1+1+5+4+6+4+1+3) \\
=\ &\tfrac{1}{10}(3\cdot 1 + 1\cdot 2 + 1\cdot 3 + 2\cdot 4 + 1\cdot 5 + 2\cdot 6) \\
=\ &\tfrac{3}{10}\cdot 1 + \tfrac{1}{10}\cdot 2 + \tfrac{1}{10}\cdot 3 + \tfrac{2}{10}\cdot 4 + \tfrac{1}{10}\cdot 5 + \tfrac{2}{10}\cdot 6 = 5.5
\end{aligned}$$

recap: the mean of a sample

- recall that the sample mean $\bar{x}$ of a dataset x is defined

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- mean of [2, 6, 1, 1, 5, 4, 6, 4, 1, 3]:

$$\begin{aligned}
&\tfrac{1}{10}(2+6+1+1+5+4+6+4+1+3) \\
&= \tfrac{1}{10}(3\cdot 1 + 1\cdot 2 + 1\cdot 3 + 2\cdot 4 + 1\cdot 5 + 2\cdot 6) \\
&= \tfrac{3}{10}\cdot 1 + \tfrac{1}{10}\cdot 2 + \tfrac{1}{10}\cdot 3 + \tfrac{2}{10}\cdot 4 + \tfrac{1}{10}\cdot 5 + \tfrac{2}{10}\cdot 6 = 5.5
\end{aligned}$$

- what happens if we roll the die many times?

the mean value of a random variable

- the notion of **mean** has a natural correspondence for random variables:

$$E(X) = \sum_i p_X(i) \cdot i$$

- this is also called the **expected value** of X
- intuitively, this corresponds to what happens if we take a very large sample from the random variable
  - and there is also a theorem called the law of large numbers that formalizes this intuition

rolling the die: mean value

- if X represents a die roll, then the mean value of X is

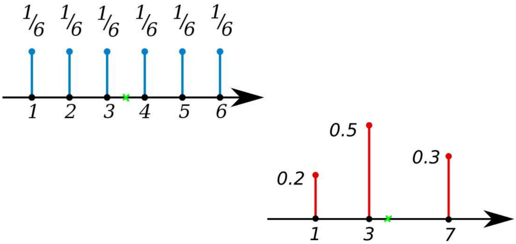
$$E(X) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3.5$$

- in general, the mean of a uniform random variable X is

$$E(X) = \frac{\text{max value} + \text{min value}}{2}$$

visual interpretation of the mean

- if we think of the pmf as weights placed on a board, $E(X)$ can be thought of as the center of mass

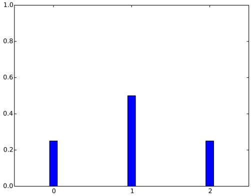


- so for all distributions with a symmetric pmf, $E(X)$ is in the middle between the lowest and the highest value

two coins: mean value

- the pmf for the number of heads when tossing two coins:

$$p_X(0) = P(X = 0) = \tfrac{1}{4}$$
$$p_X(1) = P(X = 1) = \tfrac{2}{4}$$
$$p_X(2) = P(X = 2) = \tfrac{1}{4}$$



- what's the mean?

$$\mathrm{E}(X) = \sum_i p_X(i) \cdot i = \frac{1}{4} \cdot 0 + \frac{2}{4} \cdot 1 + \frac{1}{4} \cdot 2 = 1$$

- this result makes sense – why?

UNIVERSITY OF GOTHENBURG

a few tricks with the mean

- we roll a die and multiply the result by 10; what's the mean of this r.v.?

a few tricks with the mean

- we roll a die and multiply the result by 10; what's the mean of this r.v.?
- in general:

$$\mathrm{E}(a \cdot X) = a \cdot \mathrm{E}(X)$$

a few tricks with the mean

- we roll a die and multiply the result by 10; what's the mean of this r.v.?
- in general:

$$E(a \cdot X) = a \cdot E(X)$$

- we roll two dice and sum the result; what's the mean of this r.v.?

a few tricks with the mean

- we roll a die and multiply the result by 10; what's the mean of this r.v.?
- in general:

$$E(a \cdot X) = a \cdot E(X)$$

- we roll two dice and sum the result; what's the mean of this r.v.?
- in general:

$$E(X + Y) = E(X) + E(Y)$$

variance and standard deviation

- previous lecture: the sample variance $V(x)$ of a dataset x measures how much x is concentrated to the mean
  - it is the mean of the squares of the offsets from the mean

$$V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

variance and standard deviation

- previous lecture: the sample variance $V(x)$ of a dataset x measures how much x is concentrated to the mean
  - it is the mean of the squares of the offsets from the mean

$$V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- just like for the mean value, there is a corresponding notion of **variance** for random variables: if $\mathrm{E}(X) = m$, then

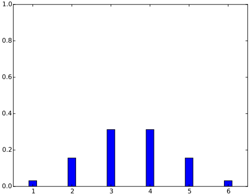
$$V(X) = \mathrm{E}[(X - m)^2]$$

- and naturally, there is also a **standard deviation**

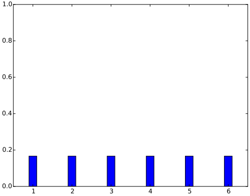
$$D(X) = \sqrt{V(X)}$$

two distributions

- low variance: pmf concentrated near the mean
  - in the extreme case: the r.v. is constant
- high variance: the pmf is more spread out



$$D(X) = 1.1$$



$$D(X) = 1.7$$

overview

recap: basic probability rules

two more basic probability rules

random variables and their distributions

mean and variance for random variables

the Bernoulli and binomial distributions

the Bernoulli and binomial distributions

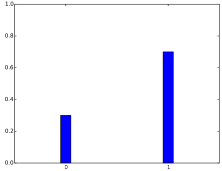
- we already saw the **uniform** distribution (die roll)
- we will have a look at two common and useful distributions:
  - **Bernoulli**: tossing an uneven coin
  - **binomial**: tossing a coin multiple times

the Bernoulli distribution

- we toss an uneven coin that gives heads ($X = 1$) with the probability p and tails ($X = 0$) with probability $1 - p$:

$$p_X(0) = 1 - p$$

$$p_X(1) = p$$

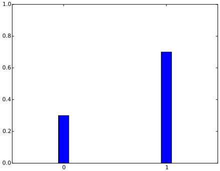


- X is then said to have a **Bernoulli** distribution with a parameter p
- this may seem like an uninteresting distribution, but it can be used as a building block for more interesting models
  - a single experiment that can "succeed" or not

the mean of the Bernoulli

- the pmf of the Bernoulli:

$$p_X(0) = 1 - p$$
$$p_X(1) = p$$

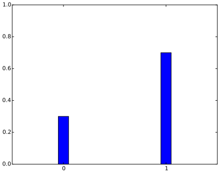


- what's the mean?

the mean of the Bernoulli

- the pmf of the Bernoulli:

$$p_X(0) = 1 - p$$
$$p_X(1) = p$$



- what's the mean?

$$\mathrm{E}(X) = \sum_i p_X(i) \cdot i = (1-p) \cdot 0 + p \cdot 1 = p$$

UNIVERSITY OF GOTHENBURG

multiple coin tosses

- we toss a coin 4 times; the probability of heads is p
- the number of heads is a r.v. X
- what is the probability of 2 heads?

multiple coin tosses

- we toss a coin 4 times; the probability of heads is p
- the number of heads is a r.v. X
- what is the probability of 2 heads?
- let's do it in two steps:
  - what's the probability of the sequence Heads-Tails-Tails-Head?

multiple coin tosses

- we toss a coin 4 times; the probability of heads is p
- the number of heads is a r.v. X
- what is the probability of 2 heads?
- let's do it in two steps:
  - what's the probability of the sequence Heads-Tails-Tails-Head?

$$P(\text{HTTH}) = p \cdot (1-p) \cdot (1-p) \cdot p = p^2 \cdot (1-p)^2$$

multiple coin tosses

- we toss a coin 4 times; the probability of heads is p
- the number of heads is a r.v. X
- what is the probability of 2 heads?
- let's do it in two steps:
  - what's the probability of the sequence Heads-Tails-Tails-Head?

$$P(\text{HTTH}) = p \cdot (1-p) \cdot (1-p) \cdot p = p^2 \cdot (1-p)^2$$

  - in how many ways can we get 2 heads?

multiple coin tosses

- we toss a coin 4 times; the probability of heads is p
- the number of heads is a r.v. X
- what is the probability of 2 heads?
- let's do it in two steps:
  - what's the probability of the sequence Heads-Tails-Tails-Head?

$$P(\text{HTTH}) = p \cdot (1-p) \cdot (1-p) \cdot p = p^2 \cdot (1-p)^2$$

  - in how many ways can we get 2 heads?

    HHTT, HTHT, HTTH, THHT, THTH, TTHH

multiple coin tosses

- we toss a coin 4 times; the probability of heads is p
- the number of heads is a r.v. X
- what is the probability of 2 heads?
- let's do it in two steps:
  - what's the probability of the sequence Heads-Tails-Tails-Head?

$$P(\text{HTTH}) = p \cdot (1-p) \cdot (1-p) \cdot p = p^2 \cdot (1-p)^2$$

  - in how many ways can we get 2 heads?

$$\text{HHTT, HTHT, HTTH, THHT, THTH, TTHH}$$

- so we get

$$P(2 \text{ heads}) = 6 \cdot p^2 \cdot (1-p)^2$$

picking k items out of n

- the number of ways to pick k items from a set of n items is called the **binomial coefficient**

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$$

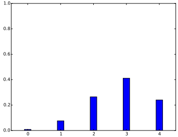
- $n! = 1 \cdot 2 \cdots n$ is the factorial function
- example: 4 coin tosses, how many combinations with k heads?

0	TTTT	1
1	HTTT, THTT, TTHT, TTTH	4
2	HHTT, HTHT, HTTH, THHT, THTH, TTHH	6
3	HHHT, HHTH, HTHH, THHH	4
4	HHHH	1

the binomial distribution

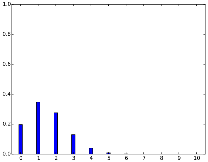
- a random variable is said to have a **binomial distribution** with parameters n and p if its pmf is

$$\binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

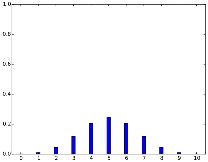


- the classical use case for the binomial distribution: **repeated experiments**
  - n corresponds to the number of experiments, p to the probability of "success"
  - this distribution will be useful when we discuss how to estimate of probabilities
- it is the sum of n independent Bernoulli

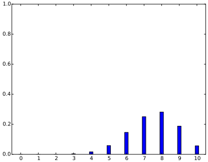
different values of p



$p = 0.15$



$p = 0.5$



$p = 0.75$

the mean of the binomial

- we toss an even coin ($p = 0.5$) 10,000 times
- roughly, how many heads do you think we get?

the mean of the binomial

- we toss an even coin ($p = 0.5$) 10,000 times
- roughly, how many heads do you think we get?
- in general, we have
$$\text{E}(X) = n \cdot p$$
- it makes sense intuitively, but can we show it theoretically?

example

- the probability that a randomly selected letter in an English word is e is 0.2
- what is the probability that an 10-letter word contains exactly three occurrences of e?

example

- the probability that a randomly selected letter in an English word is *e* is 0.2
- what is the probability that an 10-letter word contains exactly three occurrences of *e*?
  - the number of ways to put 3 *e*s into a 10-letter word, times the probability of each such word

    $$\binom{10}{3} \cdot 0.2^3 \cdot (1 - 0.2)^7 = 120 \cdot 0.2^3 \cdot 0.8^7 = 0.201$$

example

- the probability that a randomly selected letter in an English word is e is 0.2
- what is the probability that an 10-letter word contains exactly three occurrences of e?
  - the number of ways to put 3 es into a 10-letter word, times the probability of each such word

$$\binom{10}{3} \cdot 0.2^3 \cdot (1 - 0.2)^7 = 120 \cdot 0.2^3 \cdot 0.8^7 = 0.201$$

- what is the mean value of the number of occurrences of e?

example

- the probability that a randomly selected letter in an English word is e is 0.2
- what is the probability that an 10-letter word contains exactly three occurrences of e?
  - the number of ways to put 3 es into a 10-letter word, times the probability of each such word

$$\binom{10}{3} \cdot 0.2^3 \cdot (1 - 0.2)^7 = 120 \cdot 0.2^3 \cdot 0.8^7 = 0.201$$

- what is the mean value of the number of occurrences of e?

$$10 \cdot 0.2 = 2$$

next week

- on Tuesday, we'll be in the computer lab
- I'll give some more information on distributions
- first computer exercise: study distributions empirically