skipping section 6.6 / 5.6
(generating permutations and combinations)

concludes basic counting in Chapter 6 / 5

on to Chapter 7 / 6: Discrete probability
(before we go to trickier counting in Chapter 8 / 7)

Goal of Chapter 7 / 6

pages
431-475
/393-439

understanding basic probabilities,
as they pop up all over the place:

- spam filters:
  is email spam when it contains "rolex"?
- drug tests:
  are you sick when you test positive?
- evaluation of lossy channels:
  was "on" bit sent when "on" bit received?
- playing roulette/lotteries/game shows…

Introduction to discrete probability

pages 431-436 /394-398

basic definitions:

- **sample space**: a set of "possible outcomes" (hands of cards, numbers on dice)
- given a sample space S, an **experiment** results in an outcome $s \in S$
  - dealing cards (no rep.) $\Rightarrow$ hands of cards
  - rolling dice (with rep.) $\Rightarrow$ numbers
- **event**: a subset of the sample space ("three of a kind" , "sum is six")
- if S finite and each $s \in S$ is equally likely to be the result of an experiment, then **probability** of event E is $p(E) = |E|/|S|$

Complement and union of events

pages
435-436
/396-397

- event $E \subseteq S$ (sample space):
  the probability that E does not occur is
  $p(\overline{E}) = 1 - p(E)$ (use $|\overline{E}| = |S| - |E|$)
  ($\overline{E}$ is complementary event of E wrt S)

- events $E_1, E_2 \subseteq S$, then
  $p(E_1 \cup E_2) = p(E_1) + p(E_2) - p(E_1 \cap E_2)$
  proof immediate from
  $|E_1 \cup E_2| = |E_1| + |E_2| - |E_1 \cap E_2|$

- "counting" is crucial for
  elementary discrete probabilities
  (unfortunately it is not enough)

Example of discrete probabilities

pages 432-434 /394-395

urns, doors, coins, dice, cards:

- three doors, price behind only one door: probability 1/3 to win the prize
- select one card from standard deck: probability 4/52 = 1/13 it's an ace
- roll two dice: probability 5/36 that sum=6, for event = {(1,5),(2,4),(3,3),(4,2),(5,1)}

pages 436-437 /398-399 Monty & exercises

- gets complicated very easily…
- how to better model "sum of two dice"
  with sample space {2,3,…,12} and
  events with different probabilities?
- how to model unfair coin, loaded dice, …?

Probability theory, odds and ends

pages 438-452 /400-414

more flexible approach to probability needed, to deal with unfair coins, sum of dice, more contrived combinations of events, etc.

- assigning probabilities: not just $p(E) = |E|/|S|$
- conditional probability, independence
- Bernoulli trials: repeating experiments
- random variables: from outcomes to values
- birthday "paradox:" collisions unavoidable

(and, later, possibly:

- probabilistic algs: wrt time & outcome
- "the probabilistic method:" nonconstructive existence proof based on probability theory)

Assigning probabilities

pages 439-441 /401-403

to lift the $p(E) = |E|/|S|$ restriction:

let S be a countable set of outcomes

probability distribution on S is a function $p: S \rightarrow [0,1] = \mathbf{R}_{\geq 0, \leq 1}$ with $\sum_{s \in S} p(s) = 1$

thus:

- each $s \in S$ is assigned a probability $p(s)$
- for each $s \in S$: $0 \leq p(s) \leq 1$
- together ($\forall s \in S$) probabilities sum to 1: each experiment results in some outcome

define $p(E) = \sum_{s \in E} p(s)$ $\left(\leq 1 = \sum_{s \in S} p(s), \text{since } E \subseteq S\right)$

Assigning probabilities, simple remarks

pages 440-441 /402-403

- probability distribution approach covers earlier discrete probabilities: **uniform distribution** on S with $|S| = n$:

  $$\forall s \in S \quad p(s) = 1/n \quad (\Rightarrow \quad p(E) = |E|/|S|)$$

(selecting an element from a sample space with uniform distribution is sampling **at random**)

- (un)fair coin or dice, sum of dice, etc: easy to model (just make sure $\sum_{s\in S} p(s) = 1$)
- complement $p(\overline{E}) = 1 - p(E)$ and union $p(E_1 \cup E_2) = p(E_1) + p(E_2) - p(E_1 \cap E_2)$ follow as before

Conditional probability and independence

pages 441-444 /404-406

often probabilities exist in some context,
or when a certain condition is satisfied:

- what's chance to test positive
- what's chance to test positive if sick
- what's chance email is spam, if "…rolex..."

we need to be able to figure out if
context or condition influences probability:

- what's the chance of "heads" if the last five tosses were "tails"?

generally speaking: intuition cannot be trusted

Conditional probability: definition

page 442/404

let E and F be events with $p(F) > 0$

(thus $E, F \subseteq S$, for some sample space S)

the **conditional probability** of E given F

- is denoted by $p(E|F)$ (seen this in 1st semester already)
- is **defined** as $p(E \mid F) = \dfrac{p(E \cap F)}{p(F)}$
- and should be interpreted as the probability that E occurs **given the fact** that F occurs

intuition: universe S replaced by F,
event E by $E \cap F$

$\Rightarrow$ $p(E)=|E|/|S|$ by $p(E|F)=|E \cap F|/|F|=$
$(|E \cap F|/|S|)/(|F|/|S|) = p(E \cap F)/p(F)$

Conditional probability, examples

pages 442-443 /404-405

roll a die, what's probability outcome is even?

- $3/6 = ½$
- but given that outcome is ≤ 3?
  probability becomes 1/3,
  since $F = \{1,2,3\}$, $p(F) = ½$,
  $E = \{2,4,6\}$, $E\cap F=\{2\}$, $p(E\cap F)=1/6$,
  $p(E|F) = p(E\cap F)/p(F) = (1/6)/(1/2) = 1/3$

toss coin 6×; probability last toss is heads?

- ½
- but given that first five are tails?
  probability remains ½:
  $F=\{\text{tttttt},\text{ttttth}\}$, $E\cap F=\{\text{ttttth}\}$, $p(E|F)=½$

$\Rightarrow$ condition may or may not affect probability

Independence

pages 443-444 /405-406

if $p(E|F) = p(E)$, then apparently
occurrence of F does not influence E
E and F are called **independent**:

events E and F are **defined to be** independent
if $p(E \cap F) = p(E)p(F)$

($p(E|F) = p(E \cap F)/p(F) = p(E)p(F)/p(F) = p(E)$)

note that $p(F|E) = p(F)$ follows too (if $p(E) \neq 0$)

how does one decide independence?

- calculate $p(E \cap F)$, $p(E)$, and $p(F)$,
  declare independence if $p(E \cap F) = p(E)p(F)$
- in particular: don't trust your intuition

Independence examples

pages 443-444 /406

consider families with $k \geq 2$ children, assume all 2^k boy/girl configurations equally likely

- E event that family has boy(s) and girl(s)
- F event that family has at most one boy

are E and F independent?

(my) intuition useless: answer depends on k

- k=2: $p(\mathrm{E}=\{\mathrm{bg},\mathrm{gb}\})=1/2$, $p(F=\{\mathrm{bg},\mathrm{gb},\mathrm{gg}\})=3/4$
  $p(E \cap F=\{\mathrm{bg},\mathrm{gb}\})=1/2$; $p(E \cap F) \neq p(E)p(F)$: no
- k=3: $p(\mathrm{E}=\{\mathrm{bbg},\mathrm{bgb},\mathrm{bgg},\mathrm{gbb},\mathrm{gbg},\mathrm{ggb}\})=6/8$,
  $p(F=\{\mathrm{bgg},\mathrm{gbg},\mathrm{ggb},\mathrm{ggg}\})=4/8$,
  $p(E \cap F=\{\mathrm{bgg},\mathrm{gbg},\mathrm{ggb}\})=3/8 = p(E)p(F)$: yes
- k=4: … : not independent

Brief recap

pages 438-444 /400-406

- **probability distribution** on countable set of outcomes S is a function $p: S \to [0,1] = \mathbf{R}_{\geq 0, \leq 1}$ with $\sum_{s \in S} p(s) = 1$
- $E \subseteq S : p(E) = \sum_{s \in E} p(s), \quad p(\overline{E}) = 1 - p(E)$
- if $|S| = n$ and $\forall s \in S \; p(s) = 1/n$ then: **uniform distribution**, selection **at random**
- $E, F \subseteq S$
  - $p(E \cup F) = p(E) + p(F) - p(E \cap F)$
  - **conditional probability** (if $p(F) \neq 0$)

    $p(E|F) = p(E \cap F)/p(F)$
  - if $p(E \cap F) = p(E)p(F)$ then E and F are **independent** ($\leftrightarrow$ $p(E|F) = p(E)$)

Conditional probabilities, example

D: event that someone has disease d

Y: event that someone tests positive for d

events:
1. Y given D: true positive
2. Y given $\overline{D}$: false positive
3. $\overline{Y}$ given D: false negative
4. $\overline{Y}$ given $\overline{D}$: true negative

event probabilities given by lab experiments (note that 1&3 and 2&4 are complementary)

what can we say about the probability of

- D given Y (should one worry when testing positive?)
- $\overline{D}$ given $\overline{Y}$ (can one be relieved when testing negative?)

Example continued

page 456/419

suppose $p(Y|D) = 0.999$ and $p(\overline{Y} \mid \overline{D}) = 0.999$:
i.e., test is 99.9% accurate
what can we say about $p(D|Y)$ and $p(\overline{D} \mid \overline{Y})$?

generally speaking, almost nothing:
it depends on frequency of disease

if common disease, say $p(D) = 0.01$:
be concerned if test positive: $p(D|Y) > 0.9$

if rare disease, say $p(D) = 0.000001$:
be only slightly concerned if test positive:
$p(D|Y) < 0.001$
(quite a bit smaller than 0.999 …)

Based on: Bayes theorem

pages 455-457 /418-419

turning $p(Y|D)$ into $p(D|Y)$, etc:

definition: $p(Y|D) = p(Y \cap D)/p(D)$ ($p(D) \neq 0$)

$$\Rightarrow \; p(Y \cap D) = p(Y|D)p(D)$$

similarly, if $p(Y) \neq 0$: $p(Y \cap D) = p(D|Y)p(Y)$

$$\Rightarrow \; p(D|Y)p(Y) = p(Y|D)p(D)$$

$$\Rightarrow \; \boldsymbol{p(D|Y) = p(Y|D)p(D)/p(Y)}$$

with "D: have disease", "Y: test positive"

- $p(D|Y) \approx p(Y|D)$ if chances of having disease and testing positive are comparable
- $p(D|Y) << p(Y|D)$ if disease unlikely compared to testing positive for it

Bayes theorem, more useful&common form

pages 455-457 /418-419

seen that: $p(D|Y) = p(Y|D)p(D)/p(Y)$

with $Y = (Y \cap D) \cup (Y \cap \overline{D})$ a disjoint union:

$$\begin{aligned} p(Y) &= p(Y \cap D) + p(Y \cap \overline{D}) \\ &= p(Y \mid D)\, p(D) + p(Y \mid \overline{D})\, p(\overline{D}) \end{aligned}$$

Bayes thm follows (where $p(Y) \neq 0$, $p(D) \neq 0$):

$$p(D \mid Y) = \frac{p(Y \mid D)\, p(D)}{p(Y \mid D)\, p(D) + p(Y \mid \overline{D})\, p(\overline{D})}$$

Bayes theorem, details of earlier example

page 456/419

D: event to have disease d

Y: event to test positive for d

$p(Y|D) = 0.999$ and $p(\overline{Y} \mid \overline{D}) = 0.999$, thus

$p(\overline{Y} \mid D) = 0.001$ and $p(Y \mid \overline{D}) = 0.001$

If $p(D) = 0.01$:

$$p(D \mid Y) = \frac{p(Y \mid D)\, p(D)}{p(Y \mid D)\, p(D) + p(Y \mid \overline{D})\, p(\overline{D})}$$

$$= \frac{0.999 * 0.01}{0.999 * 0.01 + 0.001 * 0.99} = 0.9098$$

If $p(D) = 0.000001$:

$$p(D \mid Y) = \frac{0.999 * 0.000001}{0.999 * 0.000001 + 0.001 * 0.999999} = 0.000998$$

Bayes theorem, example details continued

if $p(D) = 0.01$:

$$p(\overline{D} \mid \overline{Y}) = \frac{p(\overline{Y} \mid \overline{D})\, p(\overline{D})}{p(\overline{Y} \mid \overline{D})\, p(\overline{D}) + p(\overline{Y} \mid D)\, p(D)}$$

$$= \frac{0.999 * 0.99}{0.999 * 0.99 + 0.001 * 0.01} > 0.999989$$

if $p(D) = 0.000001$:

$$p(\overline{D} \mid \overline{Y}) = \frac{0.999 * 0.999999}{0.999 * 0.999999 + 0.001 * 0.000001}$$

$$> 0.9999999989$$

Bayes theorem, recognizing spam

page 458/421

S: event that an email message is spam

$\overline{S}$: complementary event that it is not spam

in book: $p(S) = p(\overline{S})$:

same probabilities for spam and non-spam

more general: for instance

assume twice as much spam as non-spam

$\Rightarrow p(S) = 2/3,\ p(\overline{S}) = 1/3$

assume you've observed that

- $p(\text{"opportunity"}|S) = 1/10$
- $p(\text{"opportunity"}|\overline{S}) = 1/100$

what is probability that an email message containing "opportunity" is spam?

Bayes theorem, recognizing spam, continued

we have $p(S) = 2/3,\ p(\overline{S}) = 1/3$

and have observed that

- $p(\text{“opportunity”}|S) = 1/10$
- $p(\text{“opportunity”}|\overline{S}) = 1/100$

need the probability that an email containing “opportunity” is spam, i.e., $p(S|\text{“opportunity”})$

according to Bayes thm (w = “opportunity”):

$$p(S \mid w) = \frac{p(w \mid S)\,p(S)}{p(w \mid S)\,p(S) + p(w \mid \overline{S})\,p(\overline{S})}$$

$$= \frac{0.1 * 2/3}{0.1 * 2/3 + 0.01 * 1/3} > 0.9523$$

concludes section 7.3 / 6.3

on to expected values and variances etc, section 7.4 / 6.4

Expected values and variances

pages
463-477
/426-439

given some probability distribution:

- what can we expect to happen?
- how much fluctuation is reasonable?

intuitively: expect average over all outcomes,
each outcome weighted by its probability

- roll one die: outcomes are 1,2, …, 6,
  each with probability 1/6,
  average outcome: 1/6+2/6+…+6/6 = 3½
- roll two dice: outcomes are pairs (1,1), (1,2),
  (1,3),…,(6,6), each with probability 1/36,
  average outcome: ((1,1)+…+(6,6))/36 = ?
- tossing a coin, what's the average?

pages 446-447 /408-409

Transforming outcomes into real values: random variables

a random variable is

- not random
- not a variable

but:

- a function $S \to \mathbf{R}$, where S is a sample space

$\Rightarrow$ a random variable assigns a real value to each possible outcome in S

distribution of random variable X on S is the set of pairs $(r, p(X{=}r))$ for $r \in X(S)$, where $p(X = r) = \sum_{s \in S : X(s) = r} p(s)$

(note that this is a probability distribution)

Random variables, example

pages 446-447 /408-409

transform uniform distribution into a more general probability distribution:

let $S = \{(1,1),(1,2),\ldots,(6,6)\}$, set of outcomes of rolling two dice

define X on S by $X((i,j)) = i+j$, then:

- $X(S) = \{2,3,\ldots,12\} = S'$
- uniform distribution on S generates non-uniform probability distribution p on S':

$$p(2) = p(X = 2) = \sum_{s \in S : X(s)=2} p(s) = p((1,1)) = 1/36$$

$$p(3) = \sum_{s \in S : X(s)=3} p(s) = p((1,2)) + p((2,1) = 1/18, \text{ etc.}$$

Expected values of a random variable

pages 463-477 /426-439

given random variable X on sample space S, intuitive definition of expected value $E(X)$ becomes $E(X) = \sum_{r \in X(S)} r * p(X = r)$

with $p(X = r) = \sum_{s \in S: X(s) = r} p(s)$ it follows that

$$E(X) = \sum_{r \in X(S)} r * \left(\sum_{s \in S: X(s) = r} p(s) \right) = \sum_{s \in S} X(s) * p(s)$$

thus two different ways to compute $E(X)$:

- sum over values of X: (2/36+3/18+…+11/18+12/36 = 7)
- sum over sample space: ((2+3+3+…+11+11+12)/36 = 7)

(which one to use depends on circumstances)

Example: Bernoulli trial

pages 444-445 /406-408

an experiment with two possible outcomes (success or failure) is called a Bernoulli trial:

$\Rightarrow$ if p is success probability, then

$q = 1-p$ is the failure probability

three relevant questions:

- if same Bernoulli trial is repeated n times, what is probability of a total of k successes?
- how many successes expected after n trials?
- expect how many trials before success?

assumption: n trials **(mutually) independent**, i.e., conditional probability of success of any trial is p, conditioned on outcomes of others pages 444-445 /406-408

n independent Bernoulli trials:

if success probability of each trial is p, then the probability of precisely k successes in n independent trials is $C(n,k)p^k q^{n-k}$:

- each particular sequence of k successes (and thus $n-k$ failures) occurs with probability $p^k q^{n-k}$ (due to independence)
- there are $C(n,k)$ different sequences of k successes (sum probabilities of disjoint events)

as a function of k: the **binomial distribution** because sanity check $\sum_{k=0}^{n} C(n,k)\, p^k q^{n-k} = 1$ relies on binomial theorem

n Bernoulli trials, expected # successes

page 465/428

X: random variable counting the number of successes after n Bernoulli trials,

$\Rightarrow p(X = k) = C(n,k)p^k q^{n-k}$

$$E(X) = \sum_{k \in X(S)} k * p(X = k) \text{ where } X(S) = \{0,1,\dots,n\}$$

$$\Rightarrow \quad E(X) = \sum_{k=1}^{n} kC(n,k)\,p^k q^{n-k}$$

(pick k from n first, then leader among k, or pick leader first, then $k-1$ from $n-1$)

$$= \sum_{k=1}^{n} nC(n-1,k-1)\,p^k q^{n-k}$$

$$= \dots = np$$

(using $E(X) = \sum_{s \in S} X(s) * p(s)$: inconvenient)

n Bernoulli trials, expected # successes, easier

if X and Y are random variables on S, then

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

proof: $E(X+Y) = \sum_{s \in S} (X+Y)(s) * p(s)$

(use definition of sum of two functions)

$$= \sum_{s \in S} (X(s) + Y(s)) * p(s)$$

$$= \sum_{s \in S} X(s) * p(s) + \sum_{s \in S} Y(s) * p(s)$$

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

($E(X+Y) = \sum_{t \in (X+Y)(S)} t * p(X+Y=t)$ inconvenient)

(application: (i,j) result of two dice, $X_1((i,j))=i$, $X_2((i,j))=j$, then $E(X_1+X_2)=E(X_1)+E(X_2) = 3½+3½ = 7$)

Remaining question on Bernoulli trials

how many trials can we expect before success?

- experiment: perform trials until success
- outcomes: Y, NY, NNY, …., NN…NY, …

$\Rightarrow$ infinite sample space $S = \{Y, NY, NNY, \ldots\}$

- random variable X on S:

  $X(s)$ = number of trials needed for $s \in S$,

  thus $X(Y)=1$, $X(NY)=2$, $X(NNY)=3$, …

$\Rightarrow p(X=1)=p,\ p(X=2)=qp,\ \ldots,\ p(X=k)=q^{k-1}p, \ldots$

(note: $\sum_{k=1}^{\infty} q^{k-1} p = p \sum_{\ell=0}^{\infty} q^{\ell} = p/(1-q) = 1$,

thus called **geometric distribution**)

we need $E(X)$ (use $T(r)$, lecture 8, slide 5; $p>0$):

$$E(X) = \sum_{k=1}^{\infty} k q^{k-1} p = \ldots = p/(1-q)^2 = 1/p$$

More on expectations

pages 471-472 /434-435

seen that $E(X+Y) = E(X) + E(Y)$ for **any** random variables X and Y on sample space S

is it also true that $E(XY) = E(X)E(Y)$?

- toss coin twice, outcomes {HH,HT,TH,TT}
  X = "total number heads", so $E(X) = 1$
  Y = "total number tails", so $E(Y) = 1$
  $E(XY) = 2*0/4+1*1/4+1*1/4+0*2/4 = ½$

$\Rightarrow$ $E(XY)$ not equal to $E(X)E(Y)$

$\Rightarrow$ in general $E(XY)$ not equal to $E(X)E(Y)$

Independence of random variables

pages 471-472 /434-435

if $\forall\, x, y \in \mathbf{R}$:

$$p(X = x \text{ and } Y = y)$$

equals

$$p(X = x) * p(Y = y)$$

then X and Y are **independent**

if X and Y are independent: $E(XY) = E(X)E(Y)$

$$\begin{aligned} E(XY) &= \sum\nolimits_{r \in XY(S)} r * p(XY = r) \\ &= \sum\nolimits_{x \in X(S), y \in Y(S)} xy * p(X = x \text{ and } Y = y) \\ &= \sum\nolimits_{x \in X(S), y \in Y(S)} xy * p(X = x)\, p(Y = y) \\ &= \left(\sum\nolimits_{x \in X(S)} x * p(X = x)\right) * \left(\sum\nolimits_{y \in Y(S)} y * p(Y = y)\right) \\ &= E(X)E(Y) \end{aligned}$$

$(p(X=0 \text{ and } Y=0) = 0 \neq 1/16 = p(X=0) * p(Y=0))$

variance and standard deviation

pages 472-476 /436-439

if $\forall s \in S$: $X(s) \geq 0$ then $E(X)$ can be used to bound probability that X deviates from $E(X)$:

$$\forall x \in \mathbf{R}_{>0}: p(X \geq x) \leq E(X)/x$$

(Markov's inequality)

pf: $$E(X)/x = \sum_{r \in X(S)} (r/x)\, p(X = r)$$

$$\geq \sum_{r \in X(S), r \geq x} p(X = r) = p(X \geq x)$$

stronger result uses **variance** $V(X)$ of X,

$$V(X) = \sum_{s \in S} (X(s) - E(X))^2 p(s),$$

and the **standard deviation** $\sigma(X) = \sqrt{V(X)}$

(note: if X in "unit", then $V(X)$ in "unit^2")

Variance of a random variable

pages
472-476
/436-439

$$V(X) = \sum_{s \in S} (X(s) - E(X))^2 p(s)$$

- $V(X) = E(X^2) - (E(X))^2$ (proof: use definition)
  $\Rightarrow$ variance single Bernoulli trial is pq
- X, Y independent: $V(X+Y) = V(X) + V(Y)$
  pf.: use $V(X) = E(X^2) - (E(X))^2$,
  $E(X+Y) = E(X) + E(Y)$ (**always true**)
  and $E(XY) = E(X)E(Y)$ (due to independence)

  $\Rightarrow$ variance n indep. Bernoulli trials is npq

Chebyshev's inequality:

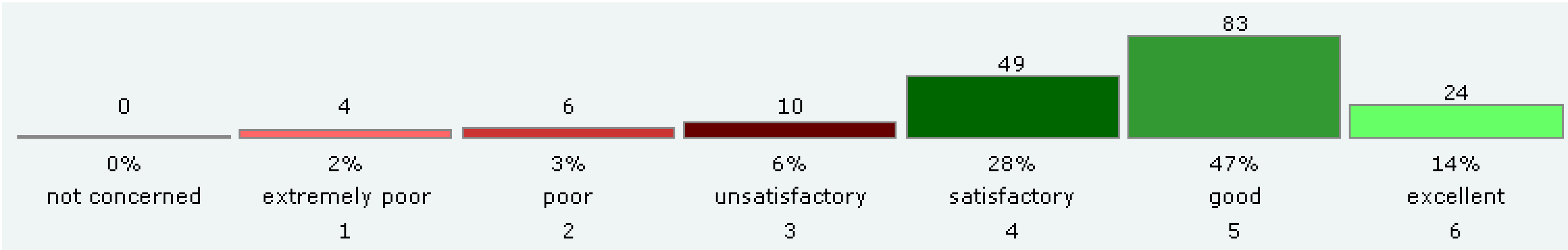
page 476/439

$$p(|X(s) - E(X)| \geq x) \leq V(X)/x^2$$

pf.: $A = \{s \in S \,|\, |X(s) - E(X)| \geq x\} \Rightarrow V(X)/x^2 \geq p(A)$

- Chebyshev's stronger than Markov's
  ($X(S) \geq 0$: $\forall x \in \mathbf{R}_{>0}\ p(X \geq x) \leq E(X)/x$)
- Chebyshev useless for $x \leq \sigma(X)$
- exponential (and non-trivial) estimates:
  use **Chernoff bound** (not here)

Variance example: course evaluation



$$E(X) = (1*4+2*6+3*10+4*49+5*83+6*24)/176 = 4.55$$

$$E(X^2) = (1^2*4+2^2*6+3^2*10+4^2*49+5^2*83+6^2*24)/176 = 21.82$$

$$\Rightarrow V(X) = E(X^2) - (E(X))^2 = 21.82 - 4.55^2 = 1.11$$

using Chebyshev's $p(|X(s) - E(X)| \geq x) \leq V(X)/x^2$, how many "extremely poor" can we expect?

$$p(|X(s) - 4.55| \geq 3.55) \leq 1.11/3.55^2 = 0.08807$$

so: at most $176*0.08807 = 15.50$

Basic probability, facts to remember

- Bayes theorem: $p(D \mid Y) = \dfrac{p(Y \mid D)\, p(D)}{p(Y \mid D)\, p(D) + p(Y \mid \overline{D})\, p(\overline{D})}$
- random variable, a function from a sample space to the real numbers
- expected value E is additive:

  for all random variables X and Y:

  $$E(X+Y) = E(X) + E(Y)$$
- variance $V(X) = E(X^2) - E(X)^2$
- for independent random variables X and Y:

  $$E(XY) = E(X)E(Y)$$

  $$V(X+Y) = V(X) + V(Y)$$
- **after about $\sqrt{n}$ drawings from n: collision**

Final remark on Ch.7/6: birthday problem

pages 447-449 /409-410

S sample space with $|S| = n$,
draw k elements at random with replacement

how likely is a **collision**,
i.e., that an element is drawn twice or more?
(applications: building hash table,
digital fingerprinting, cryptanalysis, etc.)

- if $k \leq 1$: duplicate with probability 0
- if $k > n$: duplicate with probability 1

collision probability increases with growing k,
$\Rightarrow$ for what k is collision probability $\geq$ ½ ?

purpose: show that this k is **not** about $n/2$

Birthday problem, rough analysis

pages 447-449 /409-410

drawing k random elements with replacement
from sample space S with $|S| = n$,
for what k is collision probability $\geq \frac{1}{2}$?

look at complementary problem:
analyse probability to pick k distinct elements

Probability to pick *k* distinct elements

pages
447-449
/409-410

1. if $k = 1$: probability 1 that element is unique
2. if $k = 2$: probability $1 * \frac{n-1}{n}$ to have two distinct elements
3. if $k = 3$: probability $1 * \frac{n-1}{n} * \frac{n-2}{n}$ to have three distinct elements
4. for general k: $1 * \frac{n-1}{n} * \frac{n-2}{n} * \cdots * \frac{n-k+1}{n}$ is probability to have k distinct elements

(this becomes zero for $k > n$, which is right)

Birthday problem, rough analysis continued

pages 447-449 /409-410

S sample space with $|S| = n$,
draw k elements at random with replacement

"all-distinct" probability after k drawings is

$$1 * \frac{n-1}{n} * \frac{n-2}{n} * \cdots * \frac{n-k+1}{n}$$

clearly decreasing: for what k does it get ≤ ½
(and thus collision probability ≥ ½)?

Birthday problem, rough analysis continued

pages 447-449 /409-410

for what k collision probability ≥ ½, i.e.:

$$1 * \frac{n-1}{n} * \frac{n-2}{n} * \cdots * \frac{n-k+1}{n} \leq 1/2$$

this is equivalent to

$$(n-1)(n-2)...(n-k+1) \leq n^{k-1}/2$$

hand-wavy argument:

$(n-1)(n-2)\ldots(n-k+1) = n^{k-1} - (k(k-1)/2)n^{k-2} + \ldots$

$\Rightarrow\ n^{k-1} - (k(k-1)/2)n^{k-2} + \ldots \leq n^{k-1}/2$

$\Leftrightarrow\ n^{k-1}/2 \leq (k(k-1)/2)n^{k-2} - \ldots$

$\Rightarrow$ suffices to take k a little bigger than $\sqrt{n}$

Birthday problem, conclusion

pages 447-449 /409-410

S sample space with $|S| = n$,
draw k elements at random with replacement:

after "only" $k = \sqrt{\pi n / 2}$ drawings, probability of a collision is larger than $1 - 1/e = 0.632120\ldots$

- k is lower than what intuition suggests,
  $\Rightarrow$ commonly called *birthday paradox*
- nothing paradoxical about it,
  just a consequence of $1+2+\ldots+k = k(k+1)/2$
- leads to lots of algorithms, and trouble