5/2

Mathematics for Computer Science
MIT 6.042J/18.062J

Deviation from the Mean

Example: IQ

IQ measure was constructed so that

average IQ = 100.

What fraction of the people can possibly have an IQ $\geq$ 300?

...at most 1/3

IQ Higher than 300?

If more than 1/3 have IQ $\geq$ 300, then
avg $>$ (1/3)$\cdot$300 $>$ 100 !
--a contradiction

IQ Higher than x?

In general,

$$\Pr\{IQ \geq x\} \leq \frac{100}{x}$$

IQ Higher than x?

Besides mean = 100,
we used only one fact about the distribution of IQ:
IQ is always nonnegative

Markov Bound

If R is nonnegative, then

$$\boxed{\Pr\{R \geq x\} \leq \frac{E[R]}{x}}$$

for $x > 0$

Markov Bound

- Weak
- Obvious
- Useful anyway

IQ ≥ 300, again

Suppose we are given that IQ is always ≥ 50?
Get a better bound using
(IQ - 50)
since this is now ≥ 0.

IQ ≥ 300, again

$$\Pr\{IQ \geq 300\} =$$

$$\Pr\{IQ - 50 \geq 300 - 50\}$$

$$\leq \frac{100-50}{300-50} = \frac{1}{5}$$

Improving the Markov Bound

$$\Pr\{|R-\mu| \geq x\}$$

$$= \Pr\{(R-\mu)^2 \geq x^2\}$$

by Markov:

$$\leq \frac{E[(R-\mu)^2]}{x^2}$$

variance of R

Chebyshev Bound

$$\boxed{\Pr\{|R-\mu| \geq x\} \leq \frac{\mathrm{Var}[R]}{x^{2}}}$$

$$\mathrm{Var}[R] ::= E[(R-\mu)^2]$$

$$\sigma_R ::= \sqrt{\mathrm{Var}[R]}$$

Standard Deviation

$$\boxed{\Pr\{|R-\mu| \geq x\} \leq \frac{\sigma^2}{x^2}}$$

R probably not many σ's from μ:

further than	σ	Pr ≤ 1
	2σ	Pr ≤ 1/4
	3σ	Pr ≤ 1/9
	4σ	Pr ≤ 1/16

Variance of an Indicator

I an indicator with E[I]=p:

$$\boxed{\mathrm{Var}[I]} ::= E\left[(I-p)^2\right]$$
$$= E\left[I^2\right] - 2pE\left[I\right] + p^2$$
$$= E\left[I\right] - 2p \cdot p + p^2$$
$$= p - 2p^2 + p^2 \boxed{= pq}$$

Albert R Meyer, May 2, 2011 lec 13M.37

Calculating Variance

$$\mathrm{Var}[aR+b] = a^2\,\mathrm{Var}[R]$$
$$\mathrm{Var}[R] = E[R^2] - \left(E[R]\right)^2$$

simple proofs applying linearity of E[] to the def of Var[]

Albert R Meyer, May 2, 2011 lec 13M.38

Calculating Variance

Pairwise Independent Additivity

$$\mathrm{Var}[R_1 + R_2 + \cdots + R_n]$$
$$= \mathrm{Var}[R_1] + \mathrm{Var}[R_2] + \cdots + \mathrm{Var}[R_n]$$

providing $R_1, R_2, \ldots, R_n$ are pairwise independent

again, a simple proof applying linearity of E[] to the def of Var[]

Albert R Meyer, May 2, 2011 lec 13M.46

Jacob D. Bernoulli (1659–1705)

Even the stupidest man —by some instinct of nature *per se* and by no previous instruction (this is truly amazing) —knows for sure that the more observations ...that are taken, the less the danger will be of straying from the mark.

---*Ars Conjectandi* (The Art of Guessing), 1713*

*taken from Grinstead \& Snell, http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html *Introduction to Probability*, American Mathematical Society, p. 310.

Albert R Meyer, May 2, 2011 lec 13M.48

Jacob D. Bernoulli (1659–1705)

It certainly remains to be inquired whether after the number of observations has been increased, the probability…of obtaining the true ratio…finally exceeds any given degree of certainty; or whether the problem has, so to speak, its own asymptote —that is, whether some degree of certainty is given which one can never exceed.

Albert R Meyer, May 2, 2011 lec 13M.49

Repeated Trials

Random var R with mean μ

n independent observations

$$R_1, \cdots, R_n$$

Albert R Meyer, May 2, 2011 lec 13M.57

Repeated Trials

take average:

$$A_n ::= \frac{R_1 + R_2 + \cdots + R_n}{n}$$

Bernoulli question: is it probably close to μ if n is big

$$\Pr\{|A_n - \mu| \le \delta\} = ?$$

Jacob D. Bernoulli (1659 - 1705)

Therefore, this is the problem which I now set forth and make known after I have pondered over it for twenty years. Both its novelty and its very great usefulness, coupled with its just as great difficulty, can exceed in weight and value all the remaining chapters of this thesis.

Weak Law of Large Numbers

$$\lim_{n\to\infty} \Pr\{|A_n - \mu| \le \delta\} = 1$$

$$\boxed{\lim_{n\to\infty} \Pr\{|A_n - \mu| > \delta\} = 0}$$

Weak Law of Large Numbers

will follow easily by Chebyshev & variance properties

$$\boxed{\lim_{n\to\infty} \Pr\{|A_n - \mu| > \delta\} = 0}$$

Repeated Trials

$$\boxed{E[A_n]} ::= E\left[\frac{R_1 + R_2 + \cdots + R_n}{n}\right]$$

$$= \frac{E[R_1] + E[R_2] + \cdots + E[R_n]}{n}$$

$$= \frac{n\mu}{n} \boxed{= \mu}$$

Weak Law of Large Numbers

So by Chebyshev

$$\Pr\{|A_n - \mu| > \delta\} \le \frac{\mathrm{Var}[A_n]}{\delta^2}$$

need only show

$\mathrm{Var}[A_n] \to 0$ as $n \to \infty$

Repeated Trials

$$\boxed{\mathrm{Var}\left[A_n\right]} = \mathrm{Var}\left[\frac{\left(R_1 + R_2 + \cdots + R_n\right)}{n}\right]$$

$$= \frac{\left(\mathrm{Var}\left[R_1\right] + \mathrm{Var}\left[R_2\right] + \cdots + \mathrm{Var}\left[R_n\right]\right)}{n^2}$$

$$\text{QED} = \frac{n\sigma^2}{n^2} = \boxed{\frac{\sigma^2}{n}} \quad \to 0$$

Analysis of the Proof

proof only used that $R_1,\ldots,R_n$ have

- same mean
- same variance
- & variances add
  — which follows from pairwise independence

Pairwise Independent Sampling

Theorem:
Let $R_1,\ldots,R_n$ be pairwise independent random vars with the same finite mean μ and variance σ^2. Let $A_n ::= \left(R_1 + R_2 + \cdots + R_n\right)/n$. Then

$$\boxed{\Pr\left\{\left|A_n - \mu\right| > \delta\right\} \le \frac{1}{n}\left(\frac{\sigma}{\delta}\right)^2}$$

Pairwise Independent Sampling

The punchline:
we now know how big a sample is needed to estimate the mean of any* random variable within any* desired tolerance with any* desired probability

*variance $< \infty$, tolerance > 0, probability < 1

Birthday Pairs

D ::= # pairs with matching b'days among *n* people in a *d*-day year

$$D = \sum_{1 \le i < j \le n} M_{ij}$$

M_{ij} ::= indicator that ith & jth birthdays match

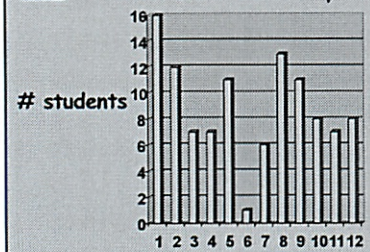
Birthday Pairs

$$E[M_{ij}] = 1/d$$

so by linearity of E[]

$$E[D] = \sum_{1 \le i < j \le n} E\left[M_{ij}\right] = \binom{n}{2} \cdot \frac{1}{d}$$

Actual Distribution by Month



Uniformity assumption not true but pretend it is

Birthday Pairs

Have data on 91 students

$$E[D] = \binom{91}{2} \cdot \frac{1}{365} \approx 11.2$$

Pairwise Independence

[Albert and Sonya have same b'day]
is independent of
[Albert and Olga have same b'day]
that is, $E_{Alice,Bob}$ & $E_{Alice,Carol}$
are independent
(pairwise, but not 3-way:
$E_{Bob,Carol}$ depends on other two)

Birthday Pairs

$\text{Var}[M_{ij}] = (1/365)(1 - 1/365)$

so by prwise linearity of Var[]

$$\text{Var}[D] = \sum \text{Var}[M_{ij}]$$

$$= \binom{91}{2} \cdot \frac{1}{365} \cdot \left(1 - \frac{1}{365}\right) \approx 11.2$$

Birthday Pairs

$$\text{Var}[M] \approx 11.1$$

$$\sigma_M < 4$$

Birthday Predictions

Chebyshev:

$\Pr\{11.2 \pm 2\sigma \text{ pairs}\} > 1 - (1/2)^2$
$= 3/4$

4 to 20 pairs 75% of the time

We actually found *exactly* 13 pairs
(& no triples)

Spring '11 Matching Birthdays

1. Jan 20
2. Jan 22
3. Jan 23
4. Apr 04
5. May 12
6. May 14
7. May 28
8. Jul 23
9. Sep 19
10. Oct 22
11. Nov 02
12. Nov 13
13. Nov 18

Team Problems

Problems 1–4

6.042 Deviations from The Mean

(4 min late)

Avg IQ = 100

What fraction can possibly have IQ ≥ 300?

- at most $\frac{1}{3}$

if avg $> (\frac{1}{3})\, 300 > 100$

Contridiction

General

$$P(IQ \geq x) \leq \frac{100}{x}$$

- make IQ a RV

(IQs can't be nonnegitive)

$$P(R \geq x) \leq \frac{E[R]}{x} \quad \text{for } x > 0$$

Markov's theorm

②

Pretty weak bound

- since very few people have IQ of 300

Obvious

but very useful anyway

Can strenghten:

Suppose told IQ is ≥ 50

So have new RV: $IQ - 50$

So $P(IQ \geq 300) = P(IQ - 50 \geq 300 - 50)$

$$\leq \frac{100 - 50}{300 - 50} = \frac{1}{5}$$

↑new, better upper bound

So get better bounds if you have a lower bound

③ Further improving Markov Bound

$$P(|R-\mu| \geq x)$$

$$= P((R-\mu)^2 \geq x^2)$$

↑
nonneg RV
can get
directly

↑ Apply Markov to this

by Markov:

$$\leq \frac{E[(R-\mu)^2]}{x^2}$$ ← Var of R

- measures how unbalanced the dist of R is

Restated as Chebyshev Bound

$$P\{|R-\mu| \geq x\} \leq \frac{Var(R)}{x^2}$$

Can increase power to ↑ accuracy

But often 4th power is ∞ on some RV which are "normal" when squared

(9)

Have to look at ∞ series to see how they behave

$$Var(R) = E[(R-\mu)^2]$$

$$\sigma_R = \sqrt{Var(R)}$$

$$var^2 = \sigma$$

St dev bounds

$$P(|R-\mu| \geq x) \leq \frac{\sigma^2}{x^2}$$

R is prob. not many σ from μ

σ $P \leq 1$

2σ $P \leq 1/4$

3σ $P \leq 1/9$

4σ $P \leq 1/16$ $\downarrow$ quadratically

Sometimes faster

- binomial $\rightarrow$ exponentially

but this is a bound

3

I is an indicator w/ $E[I] = p$

$Var(I) = \underbrace{E[(I-p)^2]}_{\text{by def}}$ by def

$= E[I^2] - 2pE[I] + p^2$ expand

$= E[I] - 2p \cdot p + p^2$ linearity of expectations

$= p - 2p^2 + p^2$

$= pq$

$q = (1-p)$

Calculating Var

$Var(aR+b) = \cancel{xa}\ a^2\ Var(R)$

multiplicative factor ^{2}

$Var(R) = E[R^2] - (E[R])^2$

⑥

Var is not linear
but easy to calc if you have pairwise ind. variables

$Var[R_1 + R_2 + \ldots R_n]$ if pairwise ind

$= Var(R_1) + Var(R_2) + \ldots + Var(R_n)$

- only dealing w/ terms that are power of 2

Story

Bernoulli came up w/ this
Said everyone could tell more obs - get closer to mean
Can you have a degree of certainty?

RV R w/ mean μ
n ind. obs
take avg $A_n = \dfrac{R_1 + R_2 + \ldots + R_n}{n}$

Bernoulli qu: Is it prob close to μ if n is big

$P(|A_n - \mu| \leq \delta) = ?$

7

Berlani took years to find this

No limit to prob

$$\lim_{n \to \infty} P(|A_n - \mu| \leq \delta) = 1$$

↑ can be as close to 1 as you want, if n is big enough, WLLN!

Follows from Chebchev + var

$$E[A_n] = E\left[\frac{R_1 + R_2 + \dots + R_n}{n}\right]$$

$$= \frac{E[R_1] + E[R_2] + \dots + E[R_n]}{n}$$

$$= \frac{n\mu}{n}$$

$$= \mu$$

8

So by Chebschev

$$P(|A_n - \mu| \geq \delta) \leq \frac{Var(A_n)}{\delta^2}$$

need only show

$Var(A_n) \to 0$ as $n \to \infty$

Repeated trials

$$Var[A_n] = Var\left[\frac{(R_1 + R_2 + \ldots + R_n)}{n}\right]$$

$$= \frac{(Var[R_1] + Var[R_2] + \ldots + Var[R_n])}{n^2}$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$ ← var of original RV r

Pcoes WLLN ?

- same mean
- " var
- their var add

9

$$P(|A_n - \mu| \geq \delta) \leq \frac{1}{n}\left(\frac{\sigma}{\delta}\right)^2$$

We now ~~kno~~ know how big a sample needs to be in order to estimate

---

Birthday Pairs

D = # matching b-days of n- people d-day year

$$= \sum_{1 \leq i \leq j \leq n} M_{ij}$$

indicator ith jth match

So $E[M_{ij}] = \frac{1}{d}$ by linearity of $E[\,]$

each persons bday

$$E[D] = \sum_{1 \leq i \leq j \leq n} E[M_{ij}] = \binom{n}{2}\frac{1}{d}$$

(10)

Birthdays are not really evenly distributed
Pretend it is uniform

$$E[D] = \binom{91}{2} \frac{1}{365} = 11.2$$

[matching birthdays

M_{ij} are pairwise ind

$$Var(M_{ij}) = \frac{1}{365} \cdot \frac{1}{365}$$

$$Var(D) = \sum var(M_{ij})$$

...

So $var(M) \approx 11.1$
is < 4

So Chebshev

$$p(11.2 \pm 20 \text{ pairs}) > 1 - \left(\frac{1}{2}\right)^2 = 3/4$$

b/w 4 and 20

[will occur $\geq \frac{3}{4}$ of the time

He found 13 matches

In-Class Problems Week 13, Mon.

Problem 1.
A herd of cows is stricken by an outbreak of *cold cow disease*. The disease lowers the normal body temperature of a cow, and a cow will die if its temperature goes below 90 degrees F. The disease epidemic is so intense that it lowered the average temperature of the herd to 85 degrees. Body temperatures as low as 70 degrees, **but no lower**, were actually found in the herd.

(a) Prove that at most 3/4 of the cows could have survived.

Hint: Let T be the temperature of a random cow. Make use of Markov's bound.

(b) Suppose there are 400 cows in the herd. Show that the bound of part (a) is best possible by giving an example set of temperatures for the cows so that the average herd temperature is 85, and with probability 3/4, a randomly chosen cow will have a high enough temperature to survive.

Problem 2.
A gambler plays 120 hands of draw poker, 60 hands of black jack, and 20 hands of stud poker per day. He wins a hand of draw poker with probability 1/6, a hand of black jack with probability 1/2, and a hand of stud poker with probability 1/5.

(a) What is the expected number of hands the gambler wins in a day?

(b) What would the Markov bound be on the probability that the gambler will win at least 108 hands on a given day? $\frac{1}{2}$

(c) Assume the outcomes of the card games are pairwise independent. What is the variance in the number of hands won per day?

(d) What would the Chebyshev bound be on the probability that the gambler will win at least 108 hands on a given day? You may answer with a numerical expression that is not completely evaluated.

Problem 3.
The proof of the Pairwise Independent Sampling Theorem 18.5.1 was given for a sequence $R_1, R_2, \ldots$ of pairwise independent random variables with the same mean and variance.

The theorem generalizes straighforwardly to sequences of pairwise independent random variables, possibly with *different* distributions, as long as all their variances are bounded by some constant.

Theorem (Generalized Pairwise Independent Sampling). *Let $X_1, X_2, \ldots$ be a sequence of pairwise independent random variables such that* $\text{Var}[X_i] \leq b$ *for some $b \geq 0$ and all $i \geq 1$. Let*

$$A_n ::= \frac{X_1 + X_2 + \cdots + X_n}{n},$$

$$\mu_n ::= \text{Ex}[A_n].$$

Then for every $\epsilon > 0$,

$$\Pr[|A_n - \mu_n| > \epsilon] \leq \frac{b}{\epsilon^2} \cdot \frac{1}{n}. \tag{1}$$

(a) Prove the Generalized Pairwise Independent Sampling Theorem.

(b) Conclude that the following holds:

Corollary (Generalized Weak Law of Large Numbers). *For every $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pr[|A_n - \mu_n| \le \epsilon] = 1.$$

Problem 4.

For any random variable, R, with mean, μ, and standard deviation, σ, the Chebyshev Bound says that for any real number $x > 0$,

$$\Pr[|R - \mu| \ge x] \le \left(\frac{\sigma}{x}\right)^2.$$

Show that for any real number, μ, and real numbers $x \ge \sigma > 0$, there is an R for which the Chebyshev Bound is tight, that is,

$$\Pr[|R| \ge x] = \left(\frac{\sigma}{x}\right)^2. \tag{2}$$

Hint: First assume $\mu = 0$ and let R only take values $0, -x$, and x.

Pairwise Independent Sampling

Let R be a random variable, and a a constant. Then

$$\operatorname{Var}[aR] = a^2 \operatorname{Var}[R]. \tag{3}$$

Theorem (Pairwise Independent Sampling). *Let $G_1, \ldots, G_n$ be pairwise independent variables with the same mean, μ, and deviation, σ. Define*

$$S_n ::= \sum_{i=1}^{n} G_i.$$

Then

$$\Pr\left[\left|\frac{S_n}{n} - \mu\right| \geq x\right] \leq \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.$$

Proof.

$$\begin{aligned}
\operatorname{Ex}\left[\frac{S_n}{n}\right] &= \operatorname{Ex}\left[\frac{\sum_{i=1}^{n} G_i}{n}\right] && \text{(def of } S_n\text{)} \\
&= \frac{\sum_{i=1}^{n} \operatorname{Ex}[G_i]}{n} && \text{(linearity of expectation)} \\
&= \frac{\sum_{i=1}^{n} \mu}{n} \\
&= \frac{n\mu}{n} = \mu.
\end{aligned}$$

$$\begin{aligned}
\operatorname{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \operatorname{Var}[S_n] && \text{(by (3))} \\
&= \frac{1}{n^2} \operatorname{Var}\left[\sum_{i=1}^{n} G_i\right] && \text{(def of } S_n\text{)} \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{Var}[G_i] && \text{(pairwise independent additivity)} \\
&= \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}. && (4)
\end{aligned}$$

This is enough to apply Chebyshev's Theorem and conclude:

$$\begin{aligned}
\Pr\left|\frac{S_n}{n} - \mu\right| \geq x &\leq \frac{\operatorname{Var}[S_n/n]}{x^2}. && \text{(Chebyshev's bound)} \\
&= \frac{\sigma^2/n}{x^2} && \text{(by (4))} \\
&= \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.
\end{aligned}$$

∎

In-Class 13 Mon

8/2

1. Cold cow disease

max $= 90°$

avg $85°$

lowest $70°$

So markov inequality w/ improvement

T = temp of random cow

$\cancel{\leq \frac{E[(R-\mu)^2]}{}}$

$$P(T-70 \geq 90-70)$$

$$\leq \frac{85-70}{90-70} = \frac{15}{20} = \frac{3}{4}$$

not really going off formula

b) 400 cows in herd

Show w/ examples w/ $P(\cancel{\ }) = \frac{3}{4}$ a randomly chosen cow will survive

(2)

Yeah that is some property I forget name of

~~B~~ But why show an example

Arrange cows so max will die

~~B~~ 300 at 90°

100 at 70° } have to split ~~bottom~~ ~~to~~ to the 2 extreams

2. 120 hands of poker $P(\text{wins} \mid \text{poker}) = \frac{1}{6}$

60 bj $\frac{1}{2}$

20 stud poker $\frac{1}{5}$

a) $E[\# \text{ wins}] = ?$

Conditional Expectations

$$E[W] = \frac{1}{6} \cdot \frac{120}{200} + \frac{1}{2} \cdot \frac{60}{200} + \frac{1}{5} \cdot \frac{20}{200}$$

$0.\bar{27}$ ↑why ?!

54

~~b)~~

↑ frac hands correct not # hands correct - trying to be too smart

(3)

b) ~~Var~~ Markov bound 108 hands

$$P(\text{wins} \geq 108) \leq \frac{54}{108}$$

$$\frac{1}{2}$$

c) Pairwise ind

Var = ?

$$E[R^2] - E[R]^2$$

$$54^2$$

$$\left(120 \cdot \frac{1}{6}\right)^2 + \left(60 \cdot \frac{1}{2}\right)^2 + \left(20 \cdot \frac{1}{5}\right)^2$$

if pairwise ind, the var just adds

Solutions to In-Class Problems Week 13, Mon.

Problem 1.
A herd of cows is stricken by an outbreak of *cold cow disease*. The disease lowers the normal body temperature of a cow, and a cow will die if its temperature goes below 90 degrees F. The disease epidemic is so intense that it lowered the average temperature of the herd to 85 degrees. Body temperatures as low as 70 degrees, **but no lower,** were actually found in the herd.

(a) Prove that at most 3/4 of the cows could have survived.

Hint: Let T be the temperature of a random cow. Make use of Markov's bound.

Solution. Let T be the temperature of a random cow. Then the fraction of cows that survive is the probability that $T \geq 90$, and $\text{Ex}[T]$ is the average temperature of the herd.

Applying Markov's Bound to T:

$$\Pr[T \geq 90] =\leq \frac{\text{Ex}[T]}{90} = \frac{85}{90} = \frac{17}{18}.$$

But $17/18 > 3/4$, so this bound is not good enough.

Instead, we apply Markov's Bound to $T - 70$:

$$\Pr[T \geq 90] = \Pr[T - 70 \geq 20] \leq \frac{\text{Ex}[T - 70]}{20} = (85 - 70)/20 = 3/4.$$

■

(b) Suppose there are 400 cows in the herd. Show that the bound of part (a) is best possible by giving an example set of temperatures for the cows so that the average herd temperature is 85, and with probability 3/4, a randomly chosen cow will have a high enough temperature to survive.

Solution. Let 100 cows have temperature 70 degrees and 300 have 90 degrees. So the probability that a random cow has a high enough temperature to survive is exactly 3/4. Also, the mean temperature is

$$(1/4)70 + (3/4)90 = 85.$$

So this distribution of temperatures satisfies the conditions under which the Markov bound implies that the probability of having a high enough temperature to survive cannot be larger than 3/4. ■

Problem 2.
A gambler plays 120 hands of draw poker, 60 hands of black jack, and 20 hands of stud poker per day. He wins a hand of draw poker with probability 1/6, a hand of black jack with probability 1/2, and a hand of stud poker with probability 1/5.

(a) What is the expected number of hands the gambler wins in a day?

Solution. $120(1/6) + 60(1/2) + 20(1/5) = 54.$ ■

(b) What would the Markov bound be on the probability that the gambler will win at least 108 hands on a given day?

Solution. The expected number of games won is 54, so by Markov, $\Pr[R \geq 108] \leq 54/108 = 1/2$. ■

(c) Assume the outcomes of the card games are pairwise independent. What is the variance in the number of hands won per day?

Solution. The variance can also be calculated using linearity of variance. For an individual hand the variance is $p(1-p)$ where p is the probability of winning. Therefore the variance is

$$120(1/6)(5/6) + 60(1/2)(1/2) + 20(1/5)(4/5) = 523/15 = 34\,\tfrac{13}{15}.$$

■

(d) What would the Chebyshev bound be on the probability that the gambler will win at least 108 hands on a given day? You may answer with a numerical expression that is not completely evaluated.

Solution.

$$\Pr[R \geq 108] = \Pr[R - 54 \geq 54] \leq \Pr[|R - 54| \geq 54] \leq \frac{\text{Var}[R]}{54^2} = \frac{523}{15(54)^2} \approx 0.01196.$$

■

Problem 3.
The proof of the Pairwise Independent Sampling Theorem 18.5.1 was given for a sequence $R_1, R_2, \ldots$ of pairwise independent random variables with the same mean and variance.

The theorem generalizes straighforwardly to sequences of pairwise independent random variables, possibly with *different* distributions, as long as all their variances are bounded by some constant.

Theorem (Generalized Pairwise Independent Sampling). *Let $X_1, X_2, \ldots$ be a sequence of pairwise independent random variables such that $\text{Var}[X_i] \leq b$ for some $b \geq 0$ and all $i \geq 1$. Let*

$$A_n ::= \frac{X_1 + X_2 + \cdots + X_n}{n},$$
$$\mu_n ::= \text{Ex}[A_n].$$

Then for every $\epsilon > 0$,

$$\Pr[|A_n - \mu_n| > \epsilon] \leq \frac{b}{\epsilon^2} \cdot \frac{1}{n}. \tag{1}$$

(a) Prove the Generalized Pairwise Independent Sampling Theorem.

Solution. Essentially identical to the proof of Theorem 18.5.1 in the text, except that G gets replaced by X and $\text{Var}[G_i]$ by b, with the equality where the b is first used becoming $\leq$. ■

(b) Conclude that the following holds:
Corollary (Generalized Weak Law of Large Numbers). *For every $\epsilon > 0$,*

$$\lim_{n \to \infty} \Pr[|A_n - \mu_n| \leq \epsilon] = 1.$$

Solution.

$$\begin{aligned}\Pr[|A_n - \mu_n| \leq \epsilon] =& 1 - \Pr[|A_n - \mu_n| > \epsilon] \\ \geq& 1 - b/(n\epsilon^2) \qquad \text{(by (1))},\end{aligned}$$

and for any fixed ϵ, this last term approaches 1 as n approaches infinity. ■

Problem 4.
For any random variable, R, with mean, μ, and standard deviation, σ, the Chebyshev Bound says that for any real number $x > 0$,

$$\Pr[|R - \mu| \geq x] \leq \left(\frac{\sigma}{x}\right)^2.$$

Show that for any real number, μ, and real numbers $x \geq \sigma > 0$, there is an R for which the Chebyshev Bound is tight, that is,

$$\Pr[|R| \geq x] = \left(\frac{\sigma}{x}\right)^2. \tag{2}$$

Hint: First assume $\mu = 0$ and let R only take values $0, -x$, and x.

Solution. From the hint, we aim to find an R with $\text{Ex}[R] = 0$ and $\text{Var}[R] = \sigma^2$ that satisfies equation (2).
Using the further hint that R takes only values $0, -x, x$, we have

$$0 = \text{Ex}[R] = x\Pr[R = x] - x\Pr[R = -x] = x\left(\Pr[R = x] - \Pr[R = -x]\right)$$

so

$$\Pr[R = x] = \Pr[R = -x], \tag{3}$$

since $x > 0$. Also,

$$\sigma^2 = \text{Ex}[R^2] = x^2\Pr[R = -x] + x^2\Pr[R = x] = 2x^2\Pr[R = x],$$

so

$$\Pr[R = x] = \frac{\sigma^2}{2x^2}.$$

This implies

$$\Pr[R = 0] = 1 - 2\Pr[R = x] = 1 - \left(\frac{\sigma}{x}\right)^2,$$

which completely determines the distribution of R. Moreover,

$$\Pr[|R| \geq x] = \Pr[R = -x] + \Pr[R = x] = 2\Pr[R = x] = \left(\frac{\sigma}{x}\right)^2$$

which confirms (2).
Finally, given μ, x, and σ, if we let $R' ::= R + \mu$, then R' will be the desired random variable for which the Chebyshev Bound is tight. ■

Pairwise Independent Sampling

Let R be a random variable, and a a constant. Then

$$\mathrm{Var}[aR] = a^2 \,\mathrm{Var}[R]. \tag{4}$$

Theorem (Pairwise Independent Sampling). *Let $G_1, \dots, G_n$ be pairwise independent variables with the same mean, μ, and deviation, σ. Define*

$$S_n ::= \sum_{i=1}^{n} G_i.$$

Then

$$\Pr[\left|\frac{S_n}{n} - \mu\right| \geq x] \leq \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.$$

Proof.

$$\begin{aligned}
\mathrm{Ex}\left[\frac{S_n}{n}\right] &= \mathrm{Ex}\left[\frac{\sum_{i=1}^{n} G_i}{n}\right] && \text{(def of } S_n\text{)}\\
&= \frac{\sum_{i=1}^{n} \mathrm{Ex}[G_i]}{n} && \text{(linearity of expectation)}\\
&= \frac{\sum_{i=1}^{n} \mu}{n}\\
&= \frac{n\mu}{n} = \mu.
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \mathrm{Var}[S_n] && \text{(by (4))}\\
&= \frac{1}{n^2}\,\mathrm{Var}\left[\sum_{i=1}^{n} G_i\right] && \text{(def of } S_n\text{)}\\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[G_i] && \text{(pairwise independent additivity)}\\
&= \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}. && (5)
\end{aligned}$$

This is enough to apply Chebyshev's Theorem and conclude:

$$\begin{aligned}
\Pr\left|\frac{S_n}{n} - \mu\right| \geq x &\leq \frac{\mathrm{Var}\,[S_n/n]}{x^2}. && \text{(Chebyshev's bound)}\\
&= \frac{\sigma^2/n}{x^2} && \text{(by (5))}\\
&= \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.
\end{aligned}$$

■

5/4

Mathematics for Computer Science
MIT 6.042J/18.062J

Sampling & Confidence

Sampling

Estimate coliform count in Charles River.
Many test stations on river.
EPA requires their average
CMD < 200:

(Coliform Microbial Density)

Sampling Questions

Choose 32 random test stations and measure CMD at each.

Sampling Questions

A few of the 32 counts turn out to be > 200 but their average is 180. Convince the EPA that avg over all stations is < 200?

Sampling Questions

That is, convince EPA that the estimate based on 32 samples is within 20 of the actual fraction?

Sampling parameters

c ::= average CMD over all stations
CMD sample $\leftrightarrow$ ran var with $\mu = c$
n stations $\leftrightarrow$ n mutually indep ran vars with $\mu = c$
A_n ::= avg CMD at the n stations

Pairwise Independent Sampling

$$\Pr\{|A_{32} - c| > 20\} \le \frac{1}{32}\left(\frac{\sigma}{20}\right)^2$$

$n = 32, \quad \mu = c, \quad \delta = 20$

?? don't know σ

Pairwise Independent Sampling

$$\Pr\{|A_{32} - c| > 20\} \le \frac{1}{32}\left(\frac{25}{20}\right)^2$$

$n = 32, \quad \mu = c, \quad \delta = 20$

worst $\sigma = \frac{L}{2}$, $L = 50$

where L is max possible difference of samples

Pairwise Independent Sampling

$$\Pr\{|A_{32} - c| > 20\} \le \frac{1}{32}\left(\frac{25}{20}\right)^2 < 0.05$$

$$\Pr\{|A_{32} - c| \le 20\} > 0.95$$

Confidence in our estimate

Tell the EPA that with probability 0.95 our estimate method for avg CMD will be within 20 of the actual avg, c, CMD over all test stations.

Confidence in our estimate

So we can be 95% confident that c, the actual avg CMD, over all stations is < 200.

Confidence —not Probable Reality

tempting to say:

"the probability that $c = 180 \pm 20$ is at least 0.95"

--technically wrong!

Confidence

c is the actual average over all test stations.

c is unknown,

but not a random variable!

 lec 13W.13

Confidence

The possible outcomes of our *sampling procedure* is a random variable. We can say that the "*probability* that our sampling process will yield an average that is $\pm$ 20 of the true average at least 0.95"

 lec 13W.14

Confidence

for simplicity we say that

$\boxed{c = 180 \pm 20 \text{ at the } 95\% \text{ confidence level}}$

 lec 13W.15

Confidence

Moral: when you are told that some fact holds at a high confidence level, remember that a random experiment lies behind this claim. Ask yourself "what experiment?"

 lec 13W.16

Team Problems

Problems 1–3

 lec 13W.17

6.042 Sampling + Confidence

5/4

- ~~NAA~~ Miniquiz 6
  - lacked the real formal reasoning
  - need to pay attention to!
  - learn the binomial stuff
  - and expanding ∞ sums

Sampling + Confidence

- Go over proof of pairwise ind

Story

Estimate coliform count in Charles

Many test stations

CMD < 200 must be for swimming

Take 32 samples of a bunch of samples
hundreds

A few are > 200

But avg = 180

Show sample ~~avg~~ (over all samples) must be < 200
hundreds

So show accurate ~~to~~ within ± 20

② C = avg CMD over all stations

its a RV - but the μ is the mean of all station

Take n samples

- mutually ind

A_n = avg CMD in the n samples

$$P(|A_n - \mu| > \delta) \le \frac{1}{n}\left(\frac{\sigma}{\delta}\right)^2$$

? what is prob that sample mean and actual mean are bigger than a threshhold δ

$n = 32$

$\mu = C$

$\delta = 20$

Want a small band that prob is small

$$P(|A_{32} - C| > 20) \le \frac{1}{32}\left(\frac{\sigma}{20}\right)^2$$

But what is σ (st. dev)?

The worst case σ is $\sigma = \frac{L}{2}$ ← max possible diff of samples $= 50/2 = 25$

? from prior history

(3)

never have stations more
Than 50 apart

So

$$P(|A_{32} - c| > 20) \le \left(\frac{25}{20}\right)^2 < .05$$

within truth

$$P(|A_{32} - c| \le 20) > .95$$

↑ 95% is special case
Can act on stuff, assume its true
95% confidence
throw people in jail

Confidence in estimate

w/ 95% Confidence our estimated ~~#~~ CMD from
32 stations is [within 20 of ~~a~~ actual
< 200 CMD

Confidence

tempting to say $\cancel{\text{prob}(C = 180 \pm 20) = .95}$ Wrong

c is reality - not an RV
- can't have a probability

Its the process / sampling procedure that is variable

* Possible outcome of sampling procedure is RV *

④

Can say "sampling process w/ yield an avg of $\pm 2\sigma$ of the true avg of at least .95"

Can say $p = 180 \pm 20$ at 95% confidence level

(prob of sampling process not prob of reality!)

"Confidence level" some random experiment lies ~~beh~~ behind the claim

- not prob (measurement is right)

In-Class Problems Week 13, Wed.

Problem 1.
A recent Gallup poll found that 35% of the adult population of the United States believes that the theory of evolution is "well-supported by the evidence." Gallup polled 1928 Americans selected uniformly and independently at random. Of these, 675 asserted belief in evolution, leading to Gallup's estimate that the fraction of Americans who believe in evolution is $675/1928 \approx 0.350$. Gallup claims a margin of error of 3 percentage points, that is, he claims to be confident that his estimate is within 0.03 of the actual percentage.

(a) What is the largest variance an indicator variable can have?

(b) Use the Pairwise Independent Sampling Theorem to determine a confidence level with which Gallup can make his claim.

(c) Gallup actually claims greater than 99% confidence in his estimate. How might he have arrived at this conclusion? (Just explain what quantity he could calculate; you do not need to carry out a calculation.)

(d) Accepting the accuracy of all of Gallup's polling data and calculations, can you conclude that there is a high probability that the number of adult Americans who believe in evolution is 35 ± 3 percent?

Problem 2.
Yesterday, the programmers at a local company wrote a large program. To estimate the fraction, b, of lines of code in this program that are buggy, the QA team will take a small sample of lines chosen randomly and independently (so it is possible, though unlikely, that the same line of code might be chosen more than once). For each line chosen, they can run tests that determine whether that line of code is buggy, after which they will use the fraction of buggy lines in their sample as their estimate of the fraction b.

The company statistician can use estimates of a binomial distribution to calculate a value, s, for a number of lines of code to sample which ensures that with 97% confidence, the fraction of buggy lines in the sample will be within 0.006 of the actual fraction, b, of buggy lines in the program.

Mathematically, the *program* is an actual outcome that already happened. The *sample* is a random variable defined by the process for randomly choosing s lines from the program. The justification for the statistician's confidence depends on some properties of the program and how the sample of s lines of code from the program are chosen. These properties are described in some of the statements below. Indicate which of these statements are true, and explain your answers.

1. The probability that the ninth line of code in the *program* is buggy is b.
2. The probability that the ninth line of code chosen for the *sample* is defective, is b.
3. All lines of code in the program are equally likely to be the third line chosen in the *sample*.
4. Given that the first line chosen for the *sample* is buggy, the probability that the second line chosen will also be buggy is greater than b.
5. Given that the last line in the *program* is buggy, the probability that the next-to-last line in the program will also be buggy is greater than b.

6. The expectation of the indicator variable for the last line in the *sample* being buggy is b.

7. Given that the first two lines of code selected in the *sample* are the same kind of statement—they might both be assignment statements, or both be conditional statements, or both loop statements,...—the probability that the first line is buggy may be greater than b.

8. There is zero probability that all the lines in the *sample* will be different.

Problem 3.
A defendent in traffic court is trying to beat a speeding ticket on the grounds that—since virtually everybody speeds on the turnpike—the police have unconstitutional discretion in giving tickets to anyone they choose. (By the way, we don't recommend this defense : -) .)

To support his argument, the defendent arranged to get a random sample of trips by 3,125 cars on the turnpike and found that 94% of them broke the speed limit at some point during their trip. He says that as a consequence of sampling theory (in particular, the Pairwise Independent Sampling Theorem), the court can be 95% confident that the actual percentage of all cars that were speeding is $94 \pm 4\%$.

The judge observes that the actual number of car trips on the turnpike was never considered in making this estimate. He is skeptical that, whether there were a thousand, a million, or 100,000,000 car trips on the turnpike, sampling only 3,125 is sufficient to be so confident.

Suppose you were were the defendent. How would you explain to the judge why the number of randomly selected cars that have to be checked for speeding *does not depend on the number of recorded trips*? Remember that judges are not trained to understand formulas, so you have to provide an intuitive, nonquantitative explanation.

Pairwise Independent Sampling

Theorem (Pairwise Independent Sampling). *Let $G_1, \ldots, G_n$ be pairwise independent variables with the same mean, μ, and deviation, σ. Define*

$$S_n ::= \sum_{i=1}^{n} G_i.$$

Then

$$\Pr[\left|\frac{S_n}{n} - \mu\right| \geq x] \leq \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.$$

In Class 13 Wed

5/4

1a. $\frac{1}{4}$ Largest σ is $\frac{1}{2}$

b) ran out of time

Solutions to In-Class Problems Week 13, Wed.

Problem 1.
A recent Gallup poll found that 35% of the adult population of the United States believes that the theory of evolution is "well-supported by the evidence." Gallup polled 1928 Americans selected uniformly and independently at random. Of these, 675 asserted belief in evolution, leading to Gallup's estimate that the fraction of Americans who believe in evolution is $675/1928 \approx 0.350$. Gallup claims a margin of error of 3 percentage points, that is, he claims to be confident that his estimate is within 0.03 of the actual percentage.

(a) What is the largest variance an indicator variable can have?

Solution.

$$\frac{1}{4}$$

By Lemma ??, $\text{Var}[H] = pq$.
Noting that $d\, p(1-p)/dp = 2p-1$ is zero when $p = 1/2$, it follows that the maximum value of $p(1-p)$ must be at $p = 1/2$, so the maximum value of $\text{Var}[H]$ is $(1/2)(1-(1/2)) = 1/4$. ■

(b) Use the Pairwise Independent Sampling Theorem to determine a confidence level with which Gallup can make his claim.

Solution. By the Pairwise Independent Sampling, the probability that a sample of size $n = 1928$ is further than $x = 0.03$ of the actual fraction is at most

$$\left(\frac{\sigma}{x}\right)^2 \cdot \frac{1}{n} \leq \left(\frac{1}{4(0.03)^2} \cdot \frac{1}{1928}\right) \leq 0.144$$

so we can be confident of Gallup's estimate at the 85.6% level. ■

(c) Gallup actually claims greater than 99% confidence in his estimate. How might he have arrived at this conclusion? (Just explain what quantity he could calculate; you do not need to carry out a calculation.)

Solution. Gallup's sample has a binomial distribution $B_{1928,p}$ for an unknown p he estimates to be about 0.35. So he wants an upper bound on

$$\Pr\left[\left|\frac{B_{1928,p}}{1928} - p\right| > 0.03\right]$$

By part (a), the variance of $B_{n,p}$ is largest when $p = 1/2$, which suggests that the probability that a sample average differs from the actual mean will be largest when $p = 1/2$. This is in fact the case. So Gallup will calculate

$$\begin{aligned}
\Pr\left[\left|\frac{B_{1928,1/2}}{1928} - \frac{1}{2}\right| > 0.03\right] &= \Pr\left[\left|B_{1928,1/2} - \frac{1928}{2}\right| > 0.03(1928)\right] \\
&= \Pr[906 \leq B_{1928,1/2} \leq 1021] \\
&= \frac{\sum_{i=906}^{1021} \binom{1928}{i}}{2^{1928}} \approx 0.9912.
\end{aligned}$$

Mathematica will actually calculate this sum exactly. There are also simple ways to use Stirling's formula to get a good estimate of this value. ■

(d) Accepting the accuracy of all of Gallup's polling data and calculations, can you conclude that there is a high probability that the number of adult Americans who believe in evolution is 35 ± 3 percent?

Solution. No. As explained in Notes and lecture, the assertion that fraction p is in the range 0.35 ± 0.03 is an assertion of fact that is either true or false. The number p is a *constant.* We don't know its value, and we don't know if the asserted fact is true or false, but there is nothing probabilistic about the fact's truth or falsehood.

We *can* say that either the assertion is true or else a 1-in-100 event occurred during the poll. Specifically, the unlikely event is that Gallup's random sample was unrepresentative. This may convince you that p is "probably" in the range 0.35 ± 0.03, but this informal "probably" is not a mathematical probability. ■

Problem 2.

Yesterday, the programmers at a local company wrote a large program. To estimate the fraction, b, of lines of code in this program that are buggy, the QA team will take a small sample of lines chosen randomly and independently (so it is possible, though unlikely, that the same line of code might be chosen more than once). For each line chosen, they can run tests that determine whether that line of code is buggy, after which they will use the fraction of buggy lines in their sample as their estimate of the fraction b.

The company statistician can use estimates of a binomial distribution to calculate a value, s, for a number of lines of code to sample which ensures that with 97% confidence, the fraction of buggy lines in the sample will be within 0.006 of the actual fraction, b, of buggy lines in the program.

Mathematically, the *program* is an actual outcome that already happened. The *sample* is a random variable defined by the process for randomly choosing s lines from the program. The justification for the statistician's confidence depends on some properties of the program and how the sample of s lines of code from the program are chosen. These properties are described in some of the statements below. Indicate which of these statements are true, and explain your answers.

1. The probability that the ninth line of code in the *program* is buggy is b.

   Solution. False.

   The program has already been written, so there's nothing probabilistic about the buggyness of the ninth (or any other) line of the program: either it is or it isn't buggy, though we don't know which. You could argue that this means it is buggy with probability zero or one, but in any case, it certainly isn't b. ■

2. The probability that the ninth line of code chosen for the *sample* is defective, is b.

   Solution. True.

   The ninth line sampled is equally likely to be any line of the program, so the probability it is buggy is the same as the fraction, b, of buggy lines in the program. ■

3. All lines of code in the program are equally likely to be the third line chosen in the *sample*.

   Solution. True.

   The meaning of "random choices of lines from the program" is precisely that at each of the s choices in the sample, in particular at the third choice, each line in the program is equally likely to be chosen. ■

4. Given that the first line chosen for the *sample* is buggy, the probability that the second line chosen will also be buggy is greater than b.

   Solution. False.

   The meaning of "*independent* random choices of lines from the program" is precisely that at each of the s choices in the sample, in particular at the second choice, each line in the program is equally likely to be chosen, independent of what the first or any other choice happened to be. ■

5. Given that the last line in the *program* is buggy, the probability that the next-to-last line in the program will also be buggy is greater than b.

   Solution. False.

   As noted above, it's zero or one. ■

6. The expectation of the indicator variable for the last line in the *sample* being buggy is b.

   Solution. True.

   The expectation of the indicator variable is the same as the probability that it is 1, namely, it is the probability that the sth line chosen is buggy, which is b, by the reasoning above. ■

7. Given that the first two lines of code selected in the *sample* are the same kind of statement—they might both be assignment statements, or both be conditional statements, or both loop statements,...—the probability that the first line is buggy may be greater than b.

   Solution. True.

   We don't know how prone to bugginess different kinds of statements may be. It could be for example, that conditionals are more prone to bugginess than other kinds of statements, and that there are more conditional lines than any other kind of line in the program. Then given that two randomly chosen lines in the sample are the same kind, they are more likely to be conditionals, which makes them more prone to bugginess. That is, the conditional probability that they will be buggy would be greater than b. ■

8. There is zero probability that all the lines in the *sample* will be different.

   Solution. False.

   We know the length, r, of the program is larger than the "small" sample size, s, in which case the probability that all the lines in the sample are different is

$$\frac{r}{r} \cdot \frac{r-1}{r} \cdot \frac{r-2}{r} \cdots \frac{r-(s-1)}{r} = \frac{r!}{(r-s)!\, r^s} > 0.$$

   Of course it would be true by the Pigeonhole Principle if $s > r$. ■

Problem 3.
A defendent in traffic court is trying to beat a speeding ticket on the grounds that—since virtually everybody speeds on the turnpike—the police have unconstitutional discretion in giving tickets to anyone they choose. (By the way, we don't recommend this defense :-).)

To support his argument, the defendent arranged to get a random sample of trips by 3,125 cars on the turnpike and found that 94% of them broke the speed limit at some point during their trip. He says that as a consequence of sampling theory (in particular, the Pairwise Independent Sampling Theorem), the court can be 95% confident that the actual percentage of all cars that were speeding is $94 \pm 4\%$.

The judge observes that the actual number of car trips on the turnpike was never considered in making this estimate. He is skeptical that, whether there were a thousand, a million, or 100,000,000 car trips on the turnpike, sampling only 3,125 is sufficient to be so confident.

Suppose you were were the defendent. How would you explain to the judge why the number of randomly selected cars that have to be checked for speeding *does not depend on the number of recorded trips*? Remember that judges are not trained to understand formulas, so you have to provide an intuitive, nonquantitative explanation.

Solution. This was intended to be a thought-provoking, conceptual question. In past terms, although most of the class could follow the derivations and crank through the formulas to calculate sample size and confidence levels, many students couldn't articulate, and indeed didn't really believe that the derived sample sizes were actually adequate to produce reliable estimates.

Here's a way to explain why we model sampling cars as independent coin tosses that might work, though we aren't sure about this.

> Of the approximately 36,000,000 recorded turnpike trips by cars in 2009, there were some *unknown* number, say 35,000,000, that broke the speed limit at some point during their trip. So in this case, the *fraction* of speeders is 35,000,000/36,000,000 which is a little over 0.97.
>
> To estimate this unknown fraction, we randomly select some trip from the 36,000,000 recorded in such a way that *every trip has an equal chance of being picked.* Picking a trip to check for speeding this way amounts to rolling a pair dice and checking that double sixes were not rolled—this has exactly the same probability as picking a speeding car.
>
> After we have picked a car trip and checked if it ever broke the speed limit, make another pick, again making sure that every recorded trip is equally likely to be picked the second time, and so on, for picking a bunch of trips. Now each pick is like rolling the dice and checking against double sixes.
>
> Now everyone understands that if we keep rolling dice looking for double sixes, then the longer we roll, the closer the fraction of rolls that are double sixes will be to 1/36, since only 1 out of the 36 possible dice ouotcomes is double six. Mathematical theory lets us calculate us how many times to roll the dice to make the fraction of double sixes very likely close to 1/36, but we needn't go into the details of the calculation.
>
> Now suppose we had a different number of recorded trips, but the same fraction were speeding. Then we could simply use the same dice in the same way to estimate the speeding fraction from this different set of trip records.
>
> So the number of rolls needed does not depend on how many trips were recorded, it just depends on the fraction of recorded speeders.

■

Pairwise Independent Sampling

Theorem (Pairwise Independent Sampling)**.** *Let $G_1, \dots, G_n$ be pairwise independent variables with the same mean, μ, and deviation, σ. Define*

$$S_n ::= \sum_{i=1}^{n} G_i.$$

Then

$$\Pr\left[\left|\frac{S_n}{n} - \mu\right| \geq x\right] \leq \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.$$

Problem Set 11

Due: May 6

Reading: Chapter 18, Deviation

Problem 1.
A coin will be flipped repeatedly until the sequence tail/tail/head (TTH) comes up. Successive flips are independent, and the coin has probability p of coming up heads. Let $N_{\mathtt{TTH}}$ be the number of coin tosses until TTH first appears. What value of p minimizes $\text{Ex}[N_{\mathtt{TTH}}]$?

Problem 2.
If R is a nonnegative random variable, then Markov's Theorem gives an upper bound on $\Pr[R \geq x]$ for any real number $x \geq \text{Ex}[R]$. If a constant $b \geq 0$ is a lower bound on R, then Markov's Theorem can also be applied to $R - b$ to obtain a possibly different bound on $\Pr[R \geq x]$.

(a) Show that if $b > 0$, applying Markov's Theorem to $R - b$ gives a smaller upper bound on $\Pr[R \geq x]$ than simply applying Markov's Theorem directly to R.

(b) What value of $b \geq 0$ in part (a) gives the best bound?

Problem 3.
The hat-check staff has had a long day serving at a party, and at the end of the party they simply return the n checked hats in a random way such that the probability that any particular person gets their own hat back is $1/n$.

Let X_i be the indicator variable for the ith person getting their own hat back. Let S_n be the total number of people who get their own hat back.

(a) What is the expected number of people who get their own hat back?

(b) Write a simple formula for $\text{Ex}[X_i X_j]$ for $i \neq j$. *Hint:* What is $\Pr\left[X_j = 1 \mid X_i = 1\right]$?

(c) Explain why you cannot use the variance of sums formula to calculate $\text{Var}[S_n]$.

(d) Show that $\text{Ex}[S_n^2] = 2$. *Hint:* $X_i^2 = X_i$.

(e) What is the variance of S_n?

(f) Show that there is at most a %1 chance that more than 10 people get their own hat back. Try to give an intuitive explanation of why the chance remains this small regardless of n.

Problem 4.
We have two coins: one is a fair coin, but the other produces heads with probability $\frac{3}{4}$. One of the two coins is picked, and this coin is tossed n times.

(a) How large must n be for you to be able to infer, with 95% confidence, which of the two coins had been chosen? (Get close to the minimum value of n required without considering any details of the relevant distribution functions, apart from mean and variance.) *Hint:* Use Chebyshev's Theorem.

(b) Suppose you had access to a computer program that would accept any $n \geq 0$ and $p \in [0, 1]$ and generate, in the form of a plot or table, the full binomial probability density and cumulative distribution functions corresponding to those parameters. How would you find the minimum number of coin flips needed to infer the identity of the chosen coin with 95% confidence? (You do not need to determine the numerical value of this minimum n, but we'd be interested to know if you did.)

Problem 5.
An *International Journal of Epidemiology* has a policy of publishing papers about drug trial results only if the conclusion about the drug's effectiveness (or lack thereof) holds at the 95% confidence level. The editors and reviewers carefully check that any trial whose results they publish was *properly performed and accurately reported.* They are also careful to check that trials whose results they publish have been conducted independently of each other.

The editors of the Journal reason that under this policy, their readership can be confident that at most 5% of the published studies will be mistaken. Later, the editors are embarrassed —and astonished —to learn that *every one* of the 20 drug trial results they published during the year was wrong. The editors thought that because the trials were conducted independently, the probability of publishing 20 wrong results was negligible, namely, $(1/20)^{20} < 10^{-25}$.

Write a brief explanation to these befuddled editors explaining what's wrong with their reasoning and how it could be that all 20 published studies were wrong.

Doing P-set 11

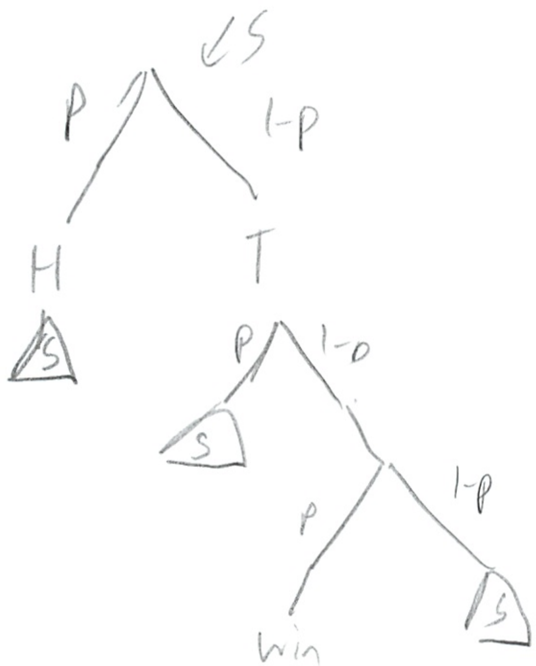
5/5

I so don't want to do
— bounds stuff

1. Coin flipped so TTH

$P = p$ heads

$N =$ TTH # tosses till this happens



Test

$S = p(1+s)$

yeah

$$S = p(1+s) + (1-p)p(2+s) + (1-p)^3(3+s)$$

(Could also do MTF where fail $=(1-p)^2 p$??

WA $$S = \frac{-3p^3 + 7p^2 - 6p + 3}{(p-1)^2 p}$$

don't think right

(2)

But MTF $\frac{1}{p}$ is also wrong

- Since can't fail on every turn, right?

Or ps must be all the same

Can't say

$1-[(1-p)^2 p] \times (1-p)^2 p$

Since # branches

Try minimizing S on WA

At p = .42525

Guess that is right

Will put it down

Makes sense

3

2. Now Makov bound.

$Q = Rv$

(non negitive)

Markov upper bound

b lower bound

a) Show if $P > 0$ gives smaller lower bound

- Was wondering about this
- dont show by example
- Show by symbols

Oh just that?

b) What value of $b \geq 0$ gives best bound?

(Someone emailed out a question on best)

What is best? try something

$P(R \geq \frac{1}{2}) \leq \frac{\mathbb{E}\frac{1}{2}}{\mathbb{E}\frac{1}{2}} = \frac{1}{2} \cdot \frac{2}{1} = 1$

$\leq$

No b

$P(R - b \geq x) \leq \frac{E[x] - b}{x}$

(4)

$$E[X] = \sum_{Y \in \text{Range } R} X\, P[X = R]$$

$$\geq \sum_{\cancel{\forall} Y \geq X} Y\, P[R = Y]$$

$$\geq X\, P[R = Y$$

$$= X\, \Sigma$$

$$= X\, P[R \geq X]$$

What is best lower bound?

No clue

Skip for now

5

3. Hat-check staff

Return n hats randomly $P(\text{get hat}) = \frac{1}{n}$

X_i = got hat back

$S_i = \sum X_i$

a) E[# got hat back) $\frac{1}{n} \cdot n$

b) Write formula for $E[X_i X_j]$ for $i \neq j$

Hint $P(X_j = 1 \mid X_i = 1)$

ind so does not matter

Needs both to be true

$\frac{1}{n} \cdot \frac{1}{n}$

Are they ind?!

- actually no!
- was in book

$$P(X_j = 1 \mid X_i = 1) = \frac{P(X_j = 1 \cap X_i = 1)}{P(X_i = 1) \leftarrow 1/n}$$

but what is this?

When I got my own hat back $n-1$ hats left

$\frac{1}{n-1}$

(4)

c) Why can't use var sum

Must be ind

d) Show $E[S_n^2] = 2$

Hint $X_i^2 = X_i$

That is S_n^2

all possible S_n^2

$1^2 = 1$

How exactly does it work

e) What is var of S_n

$Var[S_n] = E[S_n^2] - E_x^2[S_n]$

$\uparrow 2$

? ~~need to find~~

no $\left(\frac{1}{n}\right)^2$

duh

- was mixing stuff up

f) Show at most 1% chance more than 10 people get hat back.

Markov?

(7)

$$P[S_n \geq 10] \leq \frac{1/n}{10}$$

$$\leq \frac{1}{10n}$$

? not 1%

Oh n must be ≥ 10

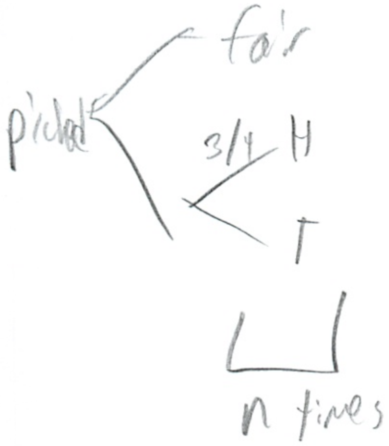
So now 1%

With more people, % gets smaller!

Note 1000 people - larger chance to get hats back

(2)

4, 2 coins



a) How large must n be for you to infer

w/ 95% what coin

Use Chebychev ← I knew that!

$$P[R - E[R] \ge x] \le \frac{Var[R]}{x^2}$$

No sampling chap

← Sum of heads

$$P\left[\left|\frac{S_n}{n} \quad - \quad\right| \ge .95\right]$$

↑ avg head

↑ B near $\frac{3}{4}$ w/ 95% confidence

Say what $\left|\frac{S_n}{n} - .75\right| \le .1$

(9)

Or am I making this too complex?

Forgot to copy a formula

$$P[|R - E[R]| \geq c\sigma_R \leq \frac{Var(R)}{(c\sigma_R)^2} = \frac{\sigma_R^2}{(c\sigma_R)^2} = \frac{1}{c^2}$$

? but must it always be $E[x]$

I think I am doing this sloppy

Don't remember anything from last time 6.041

Got something

Hope right and did not do sloppy

Why don't I get it?

- need to spend time, don't have now

b) Suppose had pc ~~game~~ program takes $n \geq 0$

$p \in [0,1]$ and makes PDF, CDF tables

How many flips needed to infer identity

How would you put this in the model

- no one single variable?
- look at mean - but that is same as above
- ? add up $0 \to .625$ $.625 \to$ etc and compare / just put something

(10)

5. Paper publishes only if confidence > 95%

Carefully check

5% study mistaken

But every one was wrong

Thought ind

? But if not ind - all use same result?

Nice only a short essay qu!

Read about pairwise ind

They say ind

- could be built on common assumption

Bad glassware

- like the moon/birth and OJ examples

(11)

Rb - ask matt

take deriv w/ respect to b

set $= 0$

$$\frac{d}{dB} \quad \frac{C + x}{d}$$

$$\frac{C}{d} + \frac{x}{d}$$

---

but is a constant!

$$P(R \geq x+b) \leq \frac{E[R]}{x+b}$$

$$\frac{d}{db}$$

$$\frac{E[R]}{(x+b)^2}$$

He doesn't know

Massachusetts Institute of Technology
6.042J/18.062J, Spring '11: Mathematics for Computer Science
Prof. Albert R Meyer

Solutions cover sheet
April 29

Student's Solutions to Problem Set 11

Your name: Michael Plasmeier

Due date: May 6

Submission date: 5/6

Circle your TA/LA: Ali Nick Oscar Oshani **Table number** 12

Collaboration statement: Circle one of the two choices and provide all pertinent info.

1. I worked alone and only with course materials.
2. I collaborated on this assignment with:

   got help from:[1] Matt Falk

   and referred to:[2]

DO NOT WRITE BELOW THIS LINE

Problem	Score
1	8
2	
3	
4	
5	
Total	

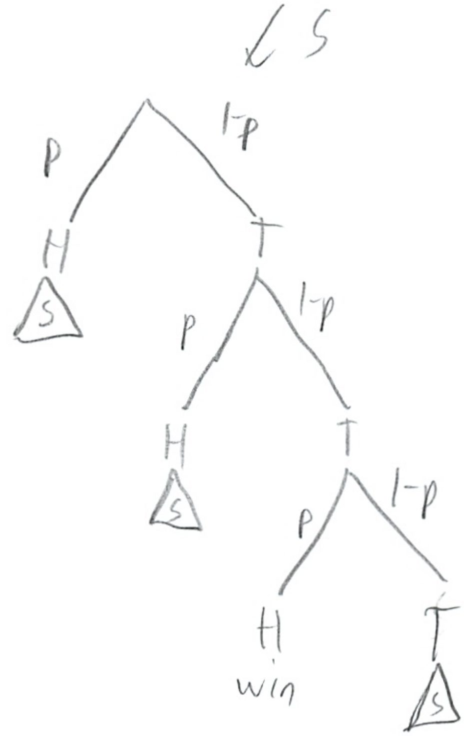
[1]People other than course staff.

[2]Give citations to texts and material other than the Spring '11 course materials.

#1: Want TTH

$p = P(\text{heads})$

$N_{TTH} = E[\#\text{ flips till } TTH]$



$S = p(1+S) + (1-p)\,p(2+S) + (1-p)^3(3+S)$

Solve for S

$$S = \frac{-3p^3 + 7p^2 - 6p + 3}{(p-1)^2\,p}$$

see solution (algebra mistake)

-2

Minimize S (Take deriv, set $= 0$)

S_{min} at $p = .42525$

Michael Plasmeier

Oshani

Table 12

P-Set 11

4

#2

a) Show $b > 0$ gives a lower bound

Without a lower bound Markov's Theorm is

$$P[R \geq x] \leq \frac{Ex[R]}{x}$$

With a lower bound b

$$P[R - b \geq x - b] \leq \frac{Ex[R] - b}{x - b}$$

? Since $b > 0$

this bound will be smaller

b) What value of b, gives best bound?

Best bound = smallest bound

Take deriv w/respect to b, set = 0

$$\cancel{\frac{\frac{d\,E[R]-b}{x}}{db}} = \frac{d}{db}\frac{E[R]}{x} - \frac{b}{x} = 0 - \frac{1}{x}$$

$$\cancel{0 = -\frac{1}{x}}$$

see solutions to see how to deal with this

②

~~$\dot{\theta} = -1/x$~~

~~False~~

"

Michael Plasmeier

Oshani

Table 12

P-Set 11

9

#3 a) E[# got own hat back] = ?

P(get own hat back) = $\frac{1}{n}$ given

So E[got own hat back] = $E[X_i] = \frac{1}{n}$ by Lemma 17.4.2

$S_n = \sum X_i = n \cdot \frac{1}{n} = 1$

b) $E[X_i X_j]$ for $i \neq j$?

? both must be 1 for $X_i X_j$ to = 1

but are not independent!

$$P[X_j = 1 \mid X_i = 1] = \frac{P[X_j = 1 \cap X_i = 1]}{P[X_i = 1]}$$

But what is this? ? 1/n

If I get my hat back, then the prob you get yours back is $\frac{1}{n-1}$

So

$$= \frac{\frac{1}{n} \frac{1}{n-1}}{\frac{1}{n}}$$

(b)

This is $\frac{1}{n-1}$

So $E[X_i X_j]$ is

$$P(X_i = 1 \cap X_j = 1) = \frac{1}{n}\frac{1}{n-1}$$

$$= \frac{1}{n(n-1)}$$

c) Why can't you use variance of sums formula?

They are not independent

d) Show that $Ex[S_n^2] = 2$

$Ex[S_n^2]$ is average of all possibe S_n^2 that is S_n squared.

S_n is the sum of X_is, Squaring these does not make a difference since $\begin{cases} 0^2 = 0 \\ 1^2 = 1 \end{cases}$

However once these numbers have been sumed, squaring <u>does</u> have an effect.

$$\begin{cases} 0^2 = 0 & \leftarrow \text{happens w/ prob } \left(1-\frac{1}{n}\right)^n \\ 1^2 = 1 & \leftarrow \text{happens w/ prob } \left(\frac{1}{n}\right)\left(1-\frac{1}{n}\right)^{n-1} \\ 2^2 = 4 & \leftarrow \text{happens w/ prob } \left(\frac{1}{n}\right)^2\left(1-\frac{1}{n}\right)^{n-2} \end{cases}$$

etc

So $E[S_n^2] = 0 \cdot \left(1-\frac{1}{n}\right)^n + 1 \cdot \left(\frac{1}{n}\right)\left(1-\frac{1}{n}\right)^{n-1} + 2\left(\frac{1}{n}\right)^2\left(1-\frac{1}{n}\right)^{n-2} + \ldots$

And the sum of this $= 2$

? by what theorem?

3

e) What is $Var(S_n)$?

$$Var(S_n) = E[S_n^2] - (E[S_n])^2$$
$$= 2 - \left(\frac{1}{n}\right)^2 \quad -1$$

f) Show at most 1% chance ≥ 10 people get their hats back, regardless of n

Markov

$$P[S_n \geq 10] \leq \frac{E[S_n]}{10}$$
$$P[S_n \geq 10] \leq \frac{1/n}{10}$$
$$\leq \frac{1}{n10}$$

$\uparrow$ $n \geq 10$ as well

So condition holds.

As $n \uparrow$, the chance of getting your hat back only falls, as we can see by the fraction.

This makes sense. If there are 1000 people in line at the coat check, then you would think 10 could get their hats back, but this is offset because the chance each person gets their own hat back is so much lower

Michael Plasmeier
Oshani
Table 12
P-Set 11

9

#4 How large must n be to know which coin w/ Confidence level 95%

$$P\Big(\Big|\underbrace{\underbrace{\underbrace{\frac{S_n}{n} - .75}_{\text{deviation from } \frac{3}{4} \text{ - if coin B}}\Big| \le .1}_{\text{should be } <.1 \text{ within } 10\%}\Big)}_{\text{That should happen 95\% of the time}} \ge .95 \quad \text{or} \quad P\Big(\Big|\frac{S_n}{n} - .5\Big| \le .1\Big) \ge .95$$

S_n ← sum of heads

n ← # of tosses

Actually can crank up the interval - don't leave a buffer

$$P\Big(\Big|\frac{S_n}{n} - .75\Big| \le .125\Big) \ge .95 \quad \text{or} \quad P\Big(\Big|\frac{S_n}{n} - .5\Big| \le .125\Big) \ge .95$$

↑ if coin B ↑ if coin A

either one will be true, telling you coin

(2)

Now need to find n

S_n is binomial distributed, so

$$Var[S_n] = n(p(1-p)) \le n \cdot \frac{1}{4} = \frac{n}{4}$$

↑ maximized when $p = 1-p$

$$Var\left[\frac{S_n}{n}\right] = \left(\frac{1}{n}\right)^2 var[S_n]$$

$$\le \left(\frac{1}{n}\right)^2 \cdot \frac{n}{4}$$

$$= \frac{1}{4n}$$

So $$P\left[\left|\frac{S_n}{n} - .75\right| \ge .125\right] \le \frac{Var[S_n/n]}{(.125)^2}$$

$$= \frac{1}{4n(.125)^2}$$

$$= \frac{16}{n}$$

And we said

$$\frac{16}{n} \le \frac{1}{20}$$

$$320 \le n$$

③

b) Suppose PC made PDF, CDF of binomial. How could you tell 95% confidence, with min n?

- My first thought would be to look at the mean - but this was part a

-? Look at the tallest bar? ×

- Look at the sum of values 1 → .625 and .625 → 1
  ↑ CDF to .625
  ? 1 - CDF to .625

- Use this as your parameter for Chebshev ✓

4

Michaël Plasmeier
Oshani
Table 12
P-Set 11

1/5

#5. The International Journal made the mistake of assuming that the studies were completly independent. Conducted independenly is a good start - but there could be some underlying connection between them. For example, they all rely on a common prior result or citation. Or they all bought bad glassware from a contaminated factory. Or the extra humidity that year messed everyone up.

statistical confidence level ≠ probability.

see solution

Solutions to Problem Set 11

Reading: Chapter **??**, Deviation

Problem 1.
A coin will be flipped repeatedly until the sequence tail/tail/head (TTH) comes up. Successive flips are independent, and the coin has probability p of coming up heads. Let N_{TTH} be the number of coin tosses until TTH first appears. What value of p minimizes $\mathrm{Ex}[N_{\mathrm{TTH}}]$?

Solution. We can describe the event tree, D, for the coin tosses as follows:

$$D = \mathrm{H} \cdot D + \mathrm{T} \cdot (\mathrm{H} \cdot D + \mathrm{T} \cdot G).$$

where

$$G = \mathrm{H} + \mathrm{T} \cdot G.$$

Abusing notation slightly, we can describe D as:

$$D = H \cdot D + T \cdot H \cdot D + T \cdot T \cdot G.$$

Now we compute $\mathrm{Ex}[N_{\mathrm{TTH}}]$:

$$\mathrm{Ex}[N_{\mathrm{TTH}}] = p\,(1 + \mathrm{Ex}[N_{\mathrm{TTH}}]) + (1-p)p\,(2 + \mathrm{Ex}[N_{\mathrm{TTH}}]) + (1-p)^2\,(2 + \mathrm{Ex}[N_{\mathrm{H}}])$$

We know $\mathrm{Ex}[N_{\mathrm{H}}] = 1/p$, so we can focus on the rest of the tree.

$$\mathrm{Ex}[N_{\mathrm{TTH}}] = (p + (1-p)p)\,\mathrm{Ex}[N_{\mathrm{TTH}}] + p + 2(1-p)p + (1-p)^2(2 + 1/p)$$

Notice how the $2 + 1/p$ term corresponds to the mean time to failure of the variable given we see TT at the start. Also, if $p = 0$ the $1/p$ term grows infinite, whereas if $p = 1$ the $(1-p)^2$ term grows infinite.

$$\begin{aligned}
&\mathrm{Ex}[N_{\mathrm{TTH}}] \\
&= \frac{p + 2(1-p)p + (1-p)^2(2 + 1/p)}{1 - p - (1-p)p} \\
&= \frac{p(3-2p)}{(1-p)^2} + 2 + \frac{1}{p}
\end{aligned}$$

Now that we have $\mathrm{Ex}[N_{\mathrm{TTH}}]$ in terms of p, we can minimize it using basic calculus. The derivative of $\mathrm{Ex}[N_{\mathrm{TTH}}]$ with respect to p is

$$\frac{1-3p}{(p-1)^3 p^2}$$

So the function is minimized at $p = 1/3$. In this case, the expected time to see TTH is $27/4 = 6\ 3/4$. Compare this to when $p = 1/2$, in that case the expected time is 8.

■

Problem 2.
If R is a nonnegative random variable, then Markov's Theorem gives an upper bound on $\Pr[R \geq x]$ for any real number $x > \text{Ex}[R]$. If a constant $b \geq 0$ is a lower bound on R, then Markov's Theorem can also be applied to $R - b$ to obtain a possibly different bound on $\Pr[R \geq x]$.

(a) Show that if $b > 0$, applying Markov's Theorem to $R - b$ gives a smaller upper bound on $\Pr[R \geq x]$ than simply applying Markov's Theorem directly to R.

Solution. Define
$$T ::= R - b.$$
Then T is a nonnegative random variable and Markov's Theorem can therefore be applied to T to give
$$\Pr[T \geq x - b] \leq \frac{\text{Ex}[T]}{x - b} = \frac{\text{Ex}[R] - b}{x - b}.$$
But the event $[T \geq x - b]$ is the same as $[R \geq x]$, so
$$\Pr[R \geq x] \leq \frac{\text{Ex}[R] - b}{x - b}.$$
So we want to show that
$$\frac{\text{Ex}[R] - b}{x - b} < \frac{\text{Ex}[R]}{x}.$$
Since x, b, and $x - b$ are all positive, therefore
$$\begin{aligned}\frac{\text{Ex}[R] - b}{x - b} &< \frac{\text{Ex}[R]}{x} \qquad \text{iff}\\ x\,\text{Ex}[R] - bx &< x\,\text{Ex}[R] - b\,\text{Ex}[R] \qquad \text{iff}\\ -bx &< -b\,\text{Ex}[R] \qquad \text{iff}\\ x &> \text{Ex}[R].\end{aligned}$$
But x *is* larger than $\text{Ex}[R]$, so
$$\frac{\text{Ex}[R] - b}{x - b} < \frac{\text{Ex}[R]}{x},$$
as required. ■

(b) What value of $b \geq 0$ in part (a) gives the best bound?

Solution. With $b \geq 0$, $R - b$ is nonnegative iff $b \in [0, \text{glb}(\text{range}(R))]$.[1] So for any such b, applying Markov's Theorem to $R - b$ gives
$$\Pr[R \geq x] \leq \frac{\text{Ex}[R] - b}{x - b}.$$
Differentiating this upper bound with respect to b gives
$$\frac{d}{db}\left(\frac{\text{Ex}[R] - b}{x - b}\right) = \frac{\text{Ex}[R] - x}{(x - b)^2}.$$
Since $x > \text{Ex}[R]$ and $x \neq b$, therefore this derivative is negative – and so the bound as a function of b is strictly decreasing – for all $b \in [0, \text{glb}(\text{range}(R))]$. Hence, the best (smallest) upper bound is given by choosing $b = \text{glb}(\text{range}(R))$.

[1] $\text{glb}(S)$ denotes the *greatest lower bound* (or *infimum*) of a set $S \subseteq \mathbb{R}$. When S is nonempty and bounded below, $\text{glb}(S)$ is just the largest real number that is no larger than any of the elements of S.

Note: To prove that the bound is strictly decreasing on the interval of interest without using calculus, let $b_1, b_2 \in [0, \text{glb}(\text{range}(R))]$. Since $x - b_1 > 0$, $x - b_2 > 0$, and $x > \text{Ex}[R]$, therefore

$$\begin{aligned}
\frac{\text{Ex}[R] - b_1}{x - b_1} < \frac{\text{Ex}[R] - b_2}{x - b_2} \qquad & \text{iff} \\
x\,\text{Ex}[R] - b_1 x - b_2\,\text{Ex}[R] + b_1 b_2 < x\,\text{Ex}[R] - b_1\,\text{Ex}[R] - b_2 x + b_1 b_2 \qquad & \text{iff} \\
(b_1 - b_2)\,\text{Ex}[R] < (b_1 - b_2) x \qquad & \text{iff} \\
b_1 - b_2 > 0 \qquad & \text{iff} \\
b_1 > b_2. &
\end{aligned}$$

■

Problem 3.
The hat-check staff has had a long day serving at a party, and at the end of the party they simply return the n checked hats in a random way such that the probability that any particular person gets their own hat back is $1/n$.

Let X_i be the indicator variable for the ith person getting their own hat back. Let S_n be the total number of people who get their own hat back.

(a) What is the expected number of people who get their own hat back?

Solution. $S_n = \sum_{i=1}^{n} X_i$, so by linearity of expectation,

$$\text{Ex}[S_n] = \sum_{1}^{n} \text{Ex}[X_i].$$

Since the probability a person gets their own hat back is $1/n$, therefore $\Pr[X_i = 1] = 1/n$. Now, since X_i is an indicator, we have $\text{Ex}[X_i] = 1/n$. By linearity of expectation,

$$\text{Ex}[S_n] = \sum_{1}^{n} \text{Ex}[X_i] = n \cdot \frac{1}{n} = 1.$$

■

(b) Write a simple formula for $\text{Ex}[X_i X_j]$ for $i \neq j$. *Hint:* What is $\Pr\left[X_j = 1 \mid X_i = 1\right]$?

Solution. We observed above that $\Pr[X_i = 1] = 1/n$. Also, given that the ith person got their own hat, each other person has an equal chance of getting their own hat among the remaining $n - 1$ hats. So

$$\Pr\left[X_j = 1 \mid X_i = 1\right] = \frac{1}{n-1},$$

for $j \neq i$. Therefore,

$$\Pr[X_i = 1 \text{ AND } X_j = 1] = \Pr\left[X_j = 1 \mid X_i = 1\right] \cdot \Pr[X_i = 1] = \frac{1}{n(n-1)}.$$

But $X_i = 1$ AND $X_j = 1$ iff $X_i X_j = 1$, so

$$\text{Ex}[X_i X_j] = \Pr[X_i X_j = 1] = \Pr[X_i = 1 \text{ AND } X_j = 1],$$

and hence

$$\text{Ex}[X_i X_j] = \frac{1}{n(n-1)}.$$

■

(c) Explain why you cannot use the variance of sums formula to calculate $\text{Var}[S_n]$.

Solution. The principle of additivity of variances requires the variables be pairwise independent, but the indicator variables for people getting their hats back are not pairwise independent, since $\Pr\left[X_j = 1 \mid X_i = 1\right] = 1/(n-1) \neq 1/n = \Pr[X_j = 1]$ for $i \neq j$. ■

(d) Show that $\text{Ex}[S_n^2] = 2$. *Hint:* $X_i^2 = X_i$.

Solution.

$$
\begin{aligned}
\text{Ex}[S_n^2] &= \text{Ex}[\sum_i X_i^2 + \sum_i \sum_{j \neq i} X_i X_j] && \text{(expanding the sum for } S_n\text{)} \\
&= \sum_i \text{Ex}[X_i^2] + \sum_i \sum_{j \neq i} \text{Ex}[X_i X_j] && \text{(linearity of Ex}[\cdot]\text{)} \\
&= \sum_i \text{Ex}[X_i] + \sum_i \sum_{j \neq i} \frac{1}{n(n-1)} && \text{(since } X_i^2 = X_i\text{)} \\
&= n \cdot \frac{1}{n} + n(n-1) \cdot \frac{1}{n(n-1)} \\
&= 2.
\end{aligned}
$$

■

(e) What is the variance of S_n?

Solution.

$$\text{Var}[S_n] = \text{Ex}[S_n^2] - \text{Ex}^2[S_n] = 2 - 1^2 = 1.$$

■

(f) Show that there is at most a 1% chance that more than 10 people get their own hat back. Try to give an intuitive explanation of why the chance remains this small regardless of n.

Solution.

$$
\begin{aligned}
\Pr[S_n \geq 11] &= \Pr[S_n - \text{Ex}[S_n] \geq 11 - \text{Ex}[S_n]] \\
&= \Pr[S_n - \text{Ex}[S_n] \geq 10] \\
&\leq \Pr[|S_n - \text{Ex}[S_n]| \geq 10] \\
&\leq \frac{\text{Var}[S_n]}{10^2} = .01
\end{aligned}
$$

TBA - intuitive explanation

■

Problem 4.

We have two coins: one is a fair coin, but the other produces heads with probability $\frac{3}{4}$. One of the two coins is picked, and this coin is tossed n times.

(a) How large must n be for you to be able to infer, with 95% confidence, which of the two coins had been chosen? (Get close to the minimum value of n required without considering any details of the relevant distribution functions, apart from mean and variance.) *Hint:* Use Chebyshev's Theorem.

Solution. To guess which coin was picked, set a threshold t between $\frac{1}{2}$ and $\frac{3}{4}$. If the proportion of heads is less than the threshold, guess that the fair coin had been picked; otherwise, guess the biased coin. Let the random variable F be the number of heads that would appear in the first n flips of the fair coin, and let B denote the number of heads that would appear in the first n flips of the biased coin. We must flip the coin sufficiently many times to ensure that

$$\Pr[\frac{F}{n} \geq t] \leq 0.05 \tag{1}$$

and

$$\Pr[\frac{B}{n} < t] \leq 0.05 \tag{2}$$

A natural threshold to choose is $t = \frac{5}{8}$, exactly in the middle of $\frac{1}{2}$ and $\frac{3}{4}$.

Now, F has an $(n, \frac{1}{2})$-binomial distribution, so its expectation and variance are $n\left(\frac{1}{2}\right) = \frac{n}{2}$ and $n\left(\frac{1}{2}\right)\left(1 - \frac{1}{2}\right) = \frac{n}{4}$, respectively. Using Chebyshev's inequality for the fair coin,

$$\begin{aligned}
\Pr[\frac{F}{n} \geq \frac{5}{8}] &= \Pr[\frac{F}{n} - \frac{1}{2} \geq \frac{5}{8} - \frac{1}{2}] = \Pr[F - \frac{n}{2} \geq \frac{n}{8}] \\
&= \Pr[F - \text{Ex}[F] \geq \frac{n}{8}] \leq \Pr[|F - \text{Ex}[F]| \geq \frac{n}{8}] \\
&\leq \frac{\text{Var}[F]}{\left(\frac{n}{8}\right)^2} = \frac{\frac{n}{4}}{\frac{n^2}{64}} = \frac{16}{n}
\end{aligned}$$

B, on the other hand, has an $(n, \frac{3}{4})$-binomial distribution, so its expectation and variance are $n\left(\frac{3}{4}\right) = \frac{3n}{4}$ and $n\left(\frac{3}{4}\right)\left(1 - \frac{3}{4}\right) = \frac{3n}{16}$, respectively. Using Chebyshev's inequality for the biased coin,

$$\begin{aligned}
\Pr[\frac{B}{n} < \frac{5}{8}] &= \Pr[\frac{3}{4} - \frac{B}{n} > \frac{3}{4} - \frac{5}{8}] = \Pr[\frac{3n}{4} - B > \frac{n}{8}] \\
&= \Pr[\text{Ex}[B] - B > \frac{n}{8}] \leq \Pr[|B - \text{Ex}[B]| \geq \frac{n}{8}] \\
&\leq \frac{\text{Var}[B]}{\left(\frac{n}{8}\right)^2} = \frac{\frac{3n}{16}}{\frac{n^2}{64}} = \frac{12}{n}
\end{aligned}$$

So, for the required confidence level, demand that $\frac{16}{n} \leq 0.05$ and $\frac{12}{n} \leq 0.05$. These hold iff $\frac{16}{n} \leq 0.05$, which is true iff $n \geq 320$. So knowing the results of at least 320 flips of the chosen coin will allow us to guess its identity with 95% confidence.

(Because the variance of the biased coin is less than that of the fair coin, we can do slightly better if we increase our threshold a bit to about 0.634, which gives 95% confidence with 279 coin flips.) ■

(b) Suppose you had access to a computer program that would accept any $n \geq 0$ and $p \in [0, 1]$ and generate, in the form of a plot or table, the full binomial probability density and cumulative distribution functions corresponding to those parameters. How would you find the minimum number of coin flips needed to infer the identity of the chosen coin with 95% confidence? (You do not need to determine the numerical value of this minimum n, but we'd be interested to know if you did.)

Solution. Again, we seek to determine the values of n that satisfy both (1) and (2). Using the same threshold as before, $t = \frac{5}{8}$, it is obvious that (1) is equivalent to

$$\mathrm{CDF}_F\left(\frac{5}{8}n\right) \geq 0.95 \tag{3}$$

while (2) is equivalent to

$$\mathrm{CDF}_B\left(\frac{5}{8}n\right) \leq 0.05 \tag{4}$$

Knowing that F is $\left(n, \frac{1}{2}\right)$-binomially distributed and B is $\left(n, \frac{3}{4}\right)$-binomially distributed, we can use the computer program to find the smallest n that satisfies both (3) and (4). ■

Problem 5.

An *International Journal of Epidemiology* has a policy of publishing papers about drug trial results only if the conclusion about the drug's effectiveness (or lack thereof) holds at the 95% confidence level. The editors and reviewers carefully check that any trial whose results they publish was *properly performed and accurately reported*. They are also careful to check that trials whose results they publish have been conducted independently of each other.

The editors of the Journal reason that under this policy, their readership can be confident that at most 5% of the published studies will be mistaken. Later, the editors are embarrassed —and astonished —to learn that *every one* of the 20 drug trial results they published during the year was wrong. The editors thought that because the trials were conducted independently, the probability of publishing 20 wrong results was negligible, namely, $(1/20)^{20} < 10^{-25}$.

Write a brief explanation to these befuddled editors explaining what's wrong with their reasoning and how it could be that all 20 published studies were wrong.

Solution. The editors have confused the statistical *confidence level* with *probability*. It's a mistake to think that because the conclusion of *particular* drug trial submitted to the journal holds at the 95% confidence level, this means its conclusion is wrong with probability only 1/20.

The conclusion of the particular submitted drug trial is right or wrong —period. An assertion of 95% confidence means that if very many trails were carried out, we expect that close to 95% of the trials would yield a correct conclusion. So if the results of all the many trials were all submitted for publication, and the editors selected 20 of these at random to publish, then they could reasonably expect that only one of them would be wrong.

But that's not what happens: not all the trials are written up and submitted. For example, there may be more than 400 worthless "alternative" drugs being tried by proponents who are genuinely honest, even if misguided. When they conduct careful trials with a 95% confidence level, we can expect that in 1/20 of the 400 trials, worthless —even damaging —drugs will look helpful. The remaining 19/20 of the 400 trials would not be submitted for publication by honest proponents because the trials did not show positive results at the 95% level. But the 20 that mistakenly showed positive results might well all be submitted with no intention to mislead.

This is why, unless there is an explanation of *why* a therapy works, scientists and doctors usually doubt results claiming to confirm the efficacy of some mysterious therapy at a high confidence level. ■

5/6

Mathematics for Computer Science
MIT 6.042J/18.062J

Avoiding Large Deviations (Chernoff Bound)

Bernoulli Sums

Focus on random vars, R, that are sums of mutually independent 0-1 variables:

$$R = \sum_{i=1}^{n} \underbrace{T_i}_{\text{Bernoulli variable}}$$

Probability of No Success

$T_i = 1$ means "success" on the i^{th} try.

$[R = 0]$ is the event that we never succeed.

Probability of No Success

Fundamental fact: Murphy's Law
If E[#successes] is large, then Pr[never succeeding] is exponentially small:

$$\Pr\left[R = 0\right] \leq e^{-\mu_R}$$

Deviation from the Mean

This is a deviation from mean result:
Pr{observed value *far from* expected value} is *SMALL*

Deviation from the Mean

$$\Pr\Big[\underbrace{R = 0}_{R \text{ below } \mu \text{ by } \underbrace{\mu}_{\text{far}}}\Big] \leq \underbrace{e^{-\mu_R}}_{\text{small (if } \mu_R \text{ large)}}$$

- only need μ_R
- don't need n
- don't need $\Pr[T_i = 1]$

Chernoff Bound

$\Pr[R \text{ above } \underbrace{c\mu}_{\text{far}}] \le \underbrace{e^{-\beta(c)\mu_R}}_{\text{small}}$

(if $c >> 1$ or μ_R large)

$$\beta(c) ::= c \cdot \ln c - c + 1$$

Chernoff Bound

still
- only need μ_R
- don't need n
- don't need $\Pr[T_i = 1]$

but

$\Pr[R \ge c\mu] \le e^{-\beta(c)\mu_R}$

Dependence on c?

c=1: $\beta(c) = 0$ (useless)

c large: $\beta(c) \approx c \log c$ LARGE

c=e: $\beta(c) = 1$

c=(1+ε): $\beta(c) = O(\varepsilon^2)$

The Lottery

Example: Pick 4

Pick a lottery number

0000, 0001, ..., 9999

The Lottery

sell 1,000,000 \$1 tickets

pay winner \$5000

μ ::= Expected #winners =

$$\frac{1{,}000{,}000}{10{,}000} = 100$$

The Lottery

sell 1,000,000 \$1 tickets

pay winner \$5000

Expected profit =

\$1,000,000 – 100 · \$5,000

= \$500,000 (good business)

The Lottery

How much reserve $$ does lottery need? Must be prepared for more than expected # winners: say a day with 60 "extra" winners?

Chernoff Bound for Lottery

Let c = 1.6, so $\beta(c) > 0.15$:

$$\Pr[R \geq \underbrace{1.6\mu}_{160}] \leq e^{-\beta(1.6)\mu}$$

Don't worry!

$= e^{-15} < 1/3{,}000{,}000$

Chance of 60 extra winners is negligible.

Large Deviation

System design must handle *rare* overloads to be reliable. That's why Chernoff more important in systems than "classical" results like the Central Limit Theorem.

Akamai Server Network

- $T_i = 1$ if ith query goes to server
- $T_i = 0$ if not
- Total Load $T = T_1 + T_2 + \ldots + T_n$
- Server averages 1M calls/day:

$$E[T] = 1{,}000{,}000$$

Designing One Server to Survive Overload

prob that rate fluctuates 1%:

$$\Pr[T \geq 1.01\ M]$$
$$\leq e^{-\beta(1.01)M}$$
$$\leq 2 \cdot 10^{-22}$$ (*very small*)!!!

1% excess capacity more than enough to make overload very unlikely.

The Whole Server Network

Akamai has a 30,000 servers, and all get same average load per day. Use Boole's inequality:

Pr [*any* server overloads]

$\leq 30{,}000 \cdot \Pr$[this server overloads]

$\leq 6 \cdot 10^{-18}$ (still very small)!!!

Chernoff vs. Binomial Bounds

If $Pr[T_i = 1]$ same for all i, then

$T = \Sigma\, T_i$ is binomial.

Get better bounds using binomial calculations, but Chernoff bound still decent.

In Class Problems

Problems 4,5

6.042 Avoiding Large Deviations

(Finish Wed 13 ~~an~~ problems)

Doctors tell you % chance you have a disease

% of people who match w/ you have condition

But misinterpeted

Usually actually statement of confidence

Or just talking about past data

- Can be misleading if you have a reason you think future will be different

Avoiding Large Deviations (Chernoff Bound)

Did Markov - initial estimate - need mean, 70

Chebchev - better - inverse square

- but need var as well
- ellusive
- can estimate
- or if you know the dist

Chernoff - need var as well

(2)

How much extra capacity do you need to avoid overloading

Have RV R

- sum of mutually ind [~~0,1~~] indicator variables

$$R = \sum_{i=1}^{n} T_i$$

↑ incator RV/Bernoulli RV

$T_i = 1 \rightarrow$ sucess on ith try

$R = 0 \rightarrow$ means never ~~happened~~ succeeded

If $E[\#\text{ sucessors}]$ is large, then

$P[R=0] \le e^{-\mu_R}$

↑ prob that nothing happens

e^{-100} is basically 0

e^{-1000} is even more 0

↑ Is <u>Murphy's Law</u>

$P[\underbrace{R=0}]$

R below μ by μ_R far if μ_R is large

③

Only need ~~μ~~ μ_R

don't need $\hat{\ }$

" " $P[T_i = 1]$

Chernov Bound

$$P[R > \underbrace{c\mu}_{\text{far}}] \le \underbrace{e^{-\beta(c)\mu_R}}_{\text{small}}$$

$\beta(c) = c \log c - c + 1$

↑ if take deriv

← is ⊕ if $c > 1$

When c is big, the $c \log c$ dominates

Dependence on c

$c = 1$ $\quad \beta(c) = 0$ useless

c large $\quad \beta(c) \approx c \log c$ Large

$c = e$ $\quad \beta(c) = 1$

$c = (1+\epsilon)$ $\quad \beta(c) = O(\epsilon^2)$

④

~~Like~~ Ex Pick 4 lotto game

1,000,000 people pay \$1 each

$\mu = E[\#\text{ of winners}] = \frac{1}{10000} \cdot 1{,}000{,}000 = 100$

Each # is $\frac{1}{10000}$ to be picked

So how much \$ does lotto need?

Pay \$5,000 to each, so \$500,000 payout

Profit \$500,000

But need a reserve fund - if 1000 winners one day?

(On lotto where pot is not split)

How likely is that?

$c = e$ so $\beta(c) = 1$

$$P[R \geq \overbrace{e\mu}^{273\text{ winners}}] \leq e^{-\mu} = e^{-100}$$

Is way small # $\approx 10^{-40}$

Safe to bet your life and need small reserve?

(5)

But can find right size of reserve ~~~~ ~~~~
given a prob reserve fund is not enough.

Systems need enough capacity to handle rare overloads,
So good estimate
- Not as good as perfect like Central Limit Theorm

Total load $T = T_1 + T_2 + \dots + T_n$
$T_i = 1$ if ith query goes to server
$T_i = 0$ if not

Server averages ~~~~ 1 M queries/day
$$E[T] = 1{,}000{,}000$$

Prob that rate will fluctiate more than 1% ?

$$P[T \geq 1.01\,M] \leq e^{-\beta(1.01)\,M}$$
$$\leq 2 \cdot 10^{-22}$$

So w/ 1% extra capacity - very negible
that it will be overloaded

⑥

Akamai has 30,000 servers

all get avg load per day

Use Boole's inequality

- prob that any one fails

= sum of prob that each one fails

$P[\text{any server fails}] \le 30000\, [\text{certain server overloads}]$

$\le 6 \cdot 10^{-18}$

So negligable

Each day

Can translate into $E[\#\text{ years before failure}]$

Prob fall off so rapidly - so can use such a coarse tool

Look at Chernoff vs Binomial Bound

if $P[T_i = 1]$ same for all i then

$R = \sum T_i$ is Binomial

Even here Chernoff bound is decent.