Overview

- Bounding A Binomial
- Web Server Load
- Sampling and Opinion Polls
- Error Correcting Codes
- Number of Pet Unicorns in Ireland

Bounding A Binomial

Suppose X is the sum of n Bernoulli random variables, $X = X_1 + X_2 + \cdots + X_n$, where random variable $X_i \sim Ber(p)$ is 1 if success in trial i and 0 otherwise.

- E.g. number of heads in n coin flips, number of corrupted bits in message sent over network
- X is a Binomial random variable: $X \sim Bin(n, p)$

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, 1, \cdots, n$$

(recall $\binom{n}{k}$ is the number of outcomes with exactly k successes and $n - k$ failures)

- Often n large e.g. $n = 12000$ bits in a 1500B packet. Often p is small e.g. bit error rate $p = 10^{-6}$.
- Then becomes hard to compute $Bin(n, p)$. Why ?
- Hint: $\binom{100}{10} \approx 10^{13}$, $\binom{100}{20}$ exceeds double precision range.

Bounding A Binomial

Extreme n and p arise commonly:

- number of errors in file written to disk
- number of elements in a particular bucket in a large hash table
- number of server crashes in a day in a large data centre
- number of facebook login requests that go to a particular server

Bounding A Binomial

Let's apply Chernoff inequality

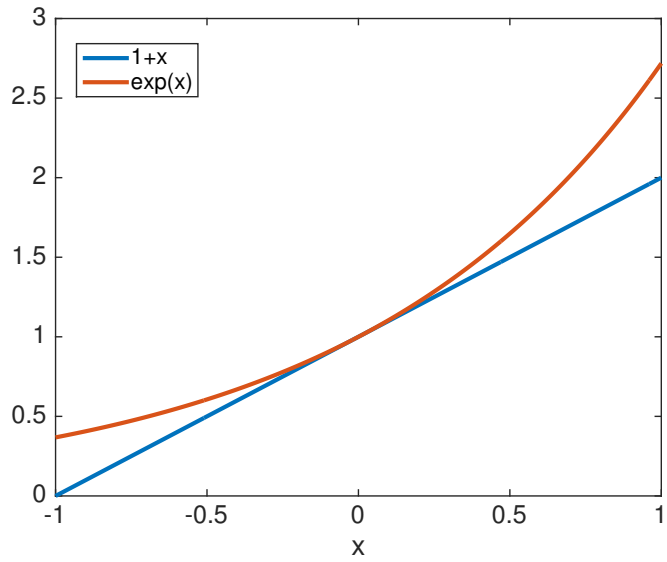
- $P(X \geq a) \leq e^{-ta} e^{\log E[e^{tX}]}$
- $X = \sum_{i=1}^{n} X_i$
- $E[e^{tX}] = E[e^{t \sum_{i=1}^{n} X_i}] = E[\prod_{i=1}^{n} e^{tX_i}]$
- Since the X_i are independent $E[\prod_{i=1}^{n} e^{tX_i}] = \prod_{i=1}^{n} E[e^{tX_i}]$
- For a single Bernoulli random variable X_i with $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$:

$$E(e^{tX_i}) = pe^t + (1-p)e^0 = pe^t + 1 - p = 1 + p(e^t - 1)$$

Bounding A Binomial

Useful trick to simplfy things:

- $E(e^{tX_i}) = 1 + p(e^t - 1) \leq e^{p(e^t - 1)}$
- Why ? $1 + x \leq e^x$



Bounding A Binomial

- So $E[e^{tX}] = \prod_{i=1}^{n}[e^{tX_i}] \leq \left(e^{p(e^t+1)}\right)^n = e^{np(e^t-1)}$
- $P(X \geq a) \leq e^{-ta}e^{\log E[e^{tX}]} \leq e^{-ta+np(e^t-1)}$
- Select $a = (1+\delta)np$

$$P(X \geq (1+\delta)np) \leq e^{-np\left(t(1+\delta)-e^t+1\right)}$$

- Try $t = \log(1+\delta)$ (Why ? Try plotting RHS vs t)

$$\begin{aligned} P(X \geq (1+\delta)np) &\leq e^{-np((1+\delta)\log(1+\delta)-(1+\delta)+1)} \\ &= e^{-np((1+\delta)\log(1+\delta)-\delta))} \end{aligned}$$

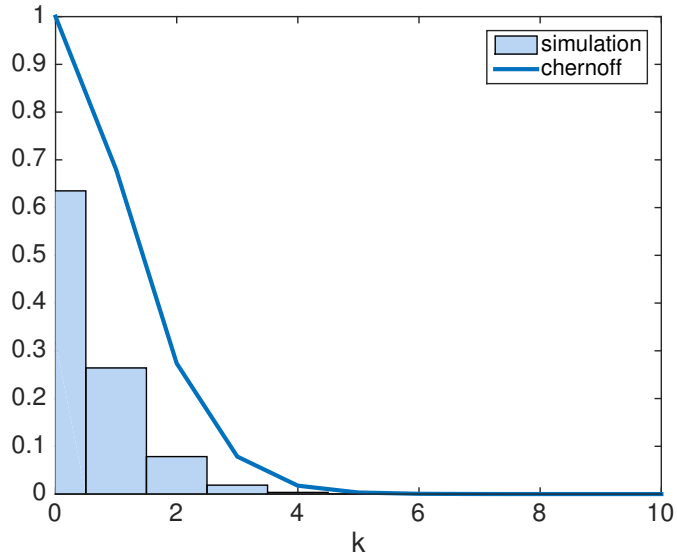
- Note that $E[X] = np$ so we can rewrite this as

$$P(X \geq (1+\delta)\mu) \leq e^{-\mu((1+\delta)\log(1+\delta)-\delta)}$$

We just need the mean μ in order to calc bound (no need for n or p) – this is v handy.

Bounding A Binomial

Let try an example[1]



$n = 100$, $p = 0.01$ (mean $np = 1$)

- Use simulation here as we can't actually evaluate the exact binomial distribution. Chernoff bound easy to evaluate, quite accurate.
- Take home message: chernoff bound v handy.

[1] p=0.01; n=100;y=[];for i=1:100000,r=rand(1,n); x=(r<=p); y=[y;sum(x)]; end; xx=[0:1:n];nn=histc(y,xx);hold off;bar(xx,1-cumsum(nn/sum(nn)),1);hold on; d=xx/(n*p);plot(xx,min(1,exp(-n*p*((1+d).*log(1+d)-d))))

Web Server Load

Requests to a web server

- Historically, server load averages 20 hits per second
- What is the probability that in 1 second we receive more than 50 hits
- Number of hits $X = \sum_i X_i$. Assume hits occur independently. Apply Chernoff bound for binomial RVs:
- $E[X] = 20 = np$. $(1+\delta)np = 50$ so $\delta = \frac{50}{np} - 1 = 2.5 - 1 = 1.5$

$$P(X \geq 30) \leq e^{-np((1+\delta)\log(1+\delta)-\delta)}$$

$$= e^{-20(2.5\log(2.5)-1.5)} \approx 10^{-7}$$

- Will almost never exceed 50 hits (assuming independence assumption valid), so enough to size server to cope with this max load.
- Why does this happen ? When we add **independent** X_i the 1's and 0's tend to cancel out provided n is large. Called **statistical multiplexing** – v important for sizing data centres, networks etc.

Sampling

Opinion poll.

- Suppose we want to know what fraction of the population likes marmite. What do you do ?
- Run a poll. Ask n people and report the fraction who like marmite.
- But how to choose n ? And how accurate is this anyway ?
- Suppose true fraction of population who likes marmite is p
- Suppose we ask n people chosen uniformly at random from the population (so we need to be careful about the way we choose people e.g. what if we only ask Australians living in Ireland ?)
- Let $X_i = 1$ if person i likes marmite and 0 otherwise. Let $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $X = nY$.
- We can use Chernoff to bound $P(X \geq (1+\delta)\mu)$ (and also $P(X \leq (1-\delta)\mu)$)

Sampling

- How do we select n so that estimate is not more than 5% above the mean 95% of the time ?
- That is, $P(Y \geq p + 0.05) = P(X \geq np + 0.05n) \leq 0.05$
- Now $P(X \geq np + 0.05n) = P(X \geq \mu + 0.05\frac{\mu}{p}) = P(X \geq (1 + \frac{0.05}{p})\mu) \leq 0.05$
- Chernoff bound tells us:

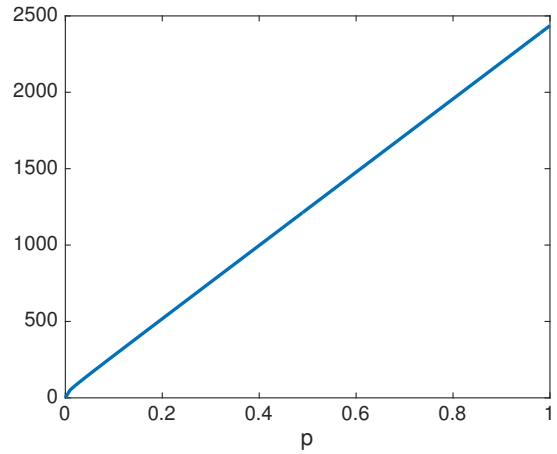
$$P(X \geq (1 + \frac{0.05}{p})\mu) \leq e^{-\mu((1+\frac{0.05}{p})\log(1+\frac{0.05}{p})-\frac{0.05}{p})}$$

- We want $e^{-\mu((1+\frac{0.05}{p})\log(1+\frac{0.05}{p})-\frac{0.05}{p})} \leq 0.05$. So need:s

$$\mu = np \geq -\frac{\log(0.05)}{(1 + \frac{0.05}{p})\log(1 + \frac{0.05}{p}) - \frac{0.05}{p}}$$

$$n \geq -\frac{\log(0.05)}{(p + 0.05)\log(1 + \frac{0.05}{p}) - 0.05}$$

Sampling



Plot of $-\frac{\log(0.05)}{(p+0.05)\log(1+\frac{0.05}{p})-0.05}$ vs p

- So we need $n \geq \approx 2436$ to ensure that 95% of the time $Y \leq p + 0.05$
- Computing a lower limit, we obtain a **confidence interval**: $p - 0.05 \leq Y \leq p + 0.05$ more than 95% of the time.

Sampling: Irish Election 2016

Election 2016: Irish Times exit poll shows Coalition well short of overall majority

Significant recovery for Fianna Fáil and gains for Sinn Féin and smaller parties, Ipsos MRBI survey finds

Fri, Feb 26, 2016, 22:48 | Updated: Sat, Feb 27, 2016, 08:05

The 2016 General Election Exit Poll was conducted exclusively on behalf of *The Irish Times* by Ipsos MRBI, among a national sample of 5,260 voters at 200 polling stations throughout all constituencies in the Republic of Ireland.

Voters were randomly selected to self-complete a mock ballot paper on exiting the polling station. The accuracy level is estimated to be approximately plus or minus 1.2 per cent.

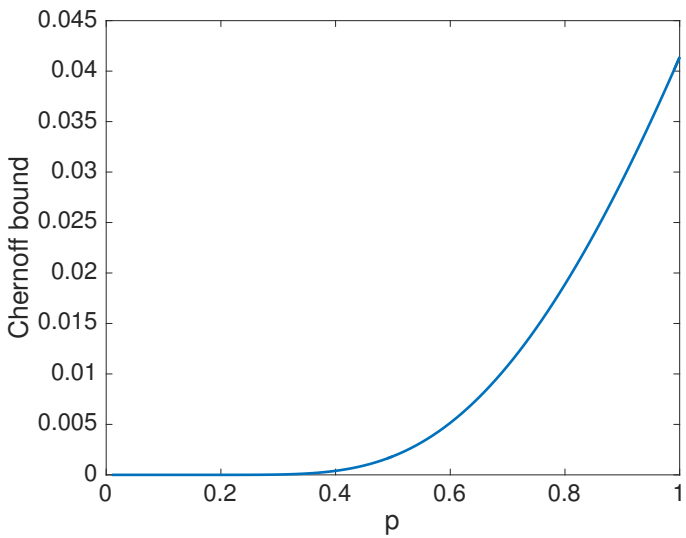
Sampling: Irish Election 2016

- RV $X_i = 1$ if person i votes for FG and 0 otherwise. Use $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$ as our estimate from sample. Let $X = nY$.
- Suppose true fraction voting FG is p, i.e. $E[Y] = p$. Chernoff:

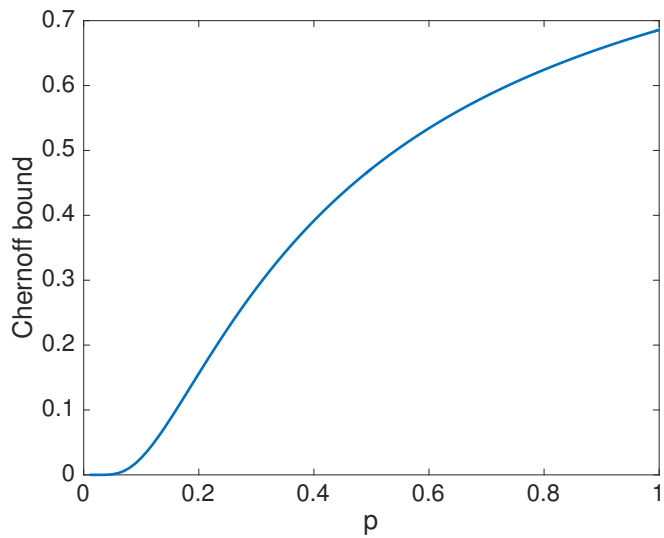
$$P(Y \geq p + \delta') = P(X \geq np + n\delta') = P(X \geq (1 + \frac{\delta'}{p})np)$$

$$P(Y \geq p + \delta') \leq e^{-np((1+\frac{\delta'}{p})\log(1+\frac{\delta'}{p})-\frac{\delta'}{p})}$$

- For $n = 5260$:



$\delta' = 0.035$

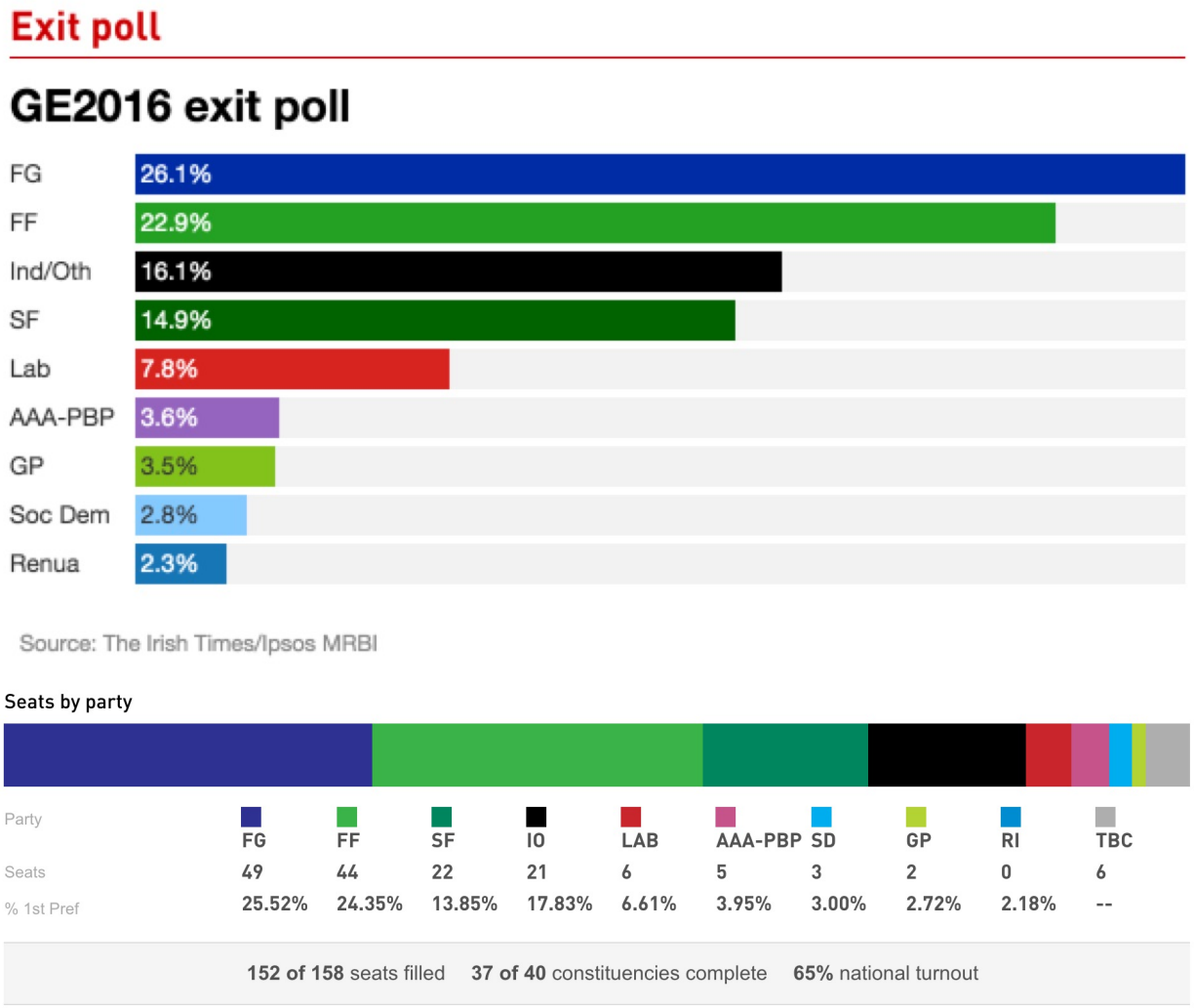


$\delta' = 0.012$

- 1.2% accuracy estimate requires extra assumptions (see CLT later).

Sampling: Irish Election 2016

How did exit poll predictions compare with final results ?



Sampling: Irish Election 2016

What might go wrong here ? We are implicitly assuming:

- Question being asked is understood $\rightarrow$ likely ok for exit poll
- Perfect response rate i.e. no refusals to respond to poll $\rightarrow$ otherwise our sample is biased.
- Honest responses $\rightarrow$ note that there is evidence of bias in exit polls[2]
- Size n is poll is fixed in advance (we don't keep polling until obtain some desired result)
- Sampling is uniformly at random without replacement $\rightarrow$ could poll be clustered ?
- Honest reporting e.g. one does not carry out multiple polls and report the best result

[2] E.g. https://en.wikipedia.org/wiki/Shy_Tory_Factor and https://en.wikipedia.org/wiki/Bradley_effect

Error Correcting Codes

- We want to transmit information packets across a lossy network. We send extra coded packets to help recover from loss.
- Suppose we send n packets of which k are information packets, so $n - k$ are coded packets.
- Ratio $r = \frac{k}{n}$ is called the **coding rate**.
- So long as we receive any k of the n transmitted packets then we can recover the k information packets.
- Suppose packet erasures are independent and identically distributed.
- Let $X_i = 1$ if packet i arrives safely, 0 otherwise. $P(X_i = 1) = 1 - p$.
- Let $X = \sum_{i=1}^{n} X_i$. $E[X] = n(1 - p)$. Probability block of packets is decoded correctly is $P(X \geq k)$.
- For a given coding rate r how big a block n do we need to use to ensure the failure probability is less than 10^{-4} ?

Error Correcting Codes

For a given coding rate r how big a block n do we need to use to ensure the failure probability is less than 10^{-4} ?

- Apply Chernoff bound:

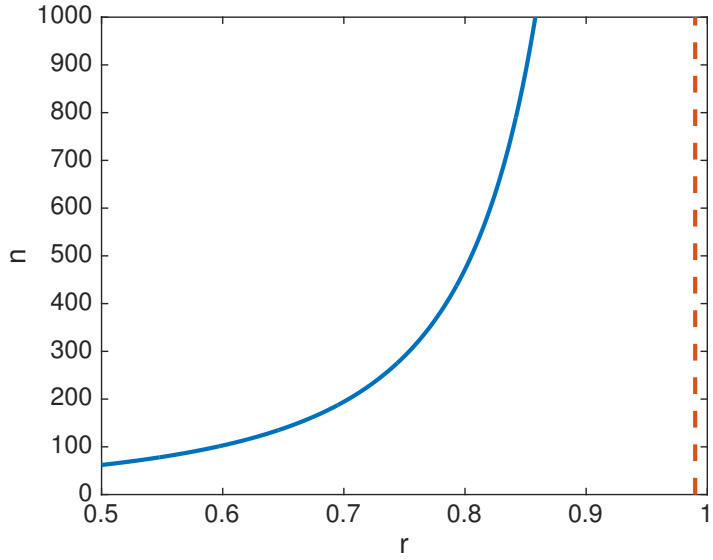
$$P(X \geq (1+\delta)\mu) \leq e^{-\mu((1+\delta)\log(1+\delta)-\delta)}$$

- Select $(1+\delta)\mu = k$, so $\delta = \frac{k}{\mu} - 1 = \frac{k}{n(1-p)} - 1 = \frac{r}{(1-p)} - 1$, $r = \frac{k}{n}$
- Select $e^{-\mu((1+\delta)\log(1+\delta)-\delta)} \leq 10^{-4}$:

$$n \geq -\frac{1}{p}\frac{\log(10^{-4})}{(1+\delta)\log(1+\delta)-\delta} = -\frac{1}{p}\frac{\log(10^{-4})}{\frac{r}{1-p}\log(\frac{r}{1-p}) - \frac{r}{1-p} + 1}$$

- Note that its the ratio $\frac{r}{1-p}$ of the coding rate to the probability that a packet is successfully received that matters

Error Correcting Codes



For $p = 0.01$ plot of $-\frac{1}{p}\frac{\log(10^{-4})}{\frac{r}{1-p}\log(\frac{r}{1-p})-\frac{r}{1-p}+1}$ vs r

- See that as r approaches $1-p$ the block size n grows v large (to infinity in fact). Why ?
- Block size is closely related to delay, since we decode a full block at a time
- For small-ish block sizes we need to use a low coding rate i.e. sacrifice throughput (send a larger fraction of coded packets)

Number of Pet Unicorns in Ireland

Opinions polls again. Suppose we carry out a survey asking $N = 1000$ people in Ireland whether they own a pet unicorn.

- Random variable $X_i = 1$ if say own a unicorn and 0 otherwise
- Most people answer "no" ($X_i = 0$).
- Suppose 1 person says "yes" ($X_i = 1$).
- So we observe that $\sum_{i=1}^{N} X_i = 1$
- Say we now try to estimate the number of pet unicorns using:

$$\left(\sum_{i=1}^{N} X_i \right) \times \frac{\text{population of Ireland}}{N}$$

Population of Ireland is about 4.5M, so we estimate $1 \times \frac{4.5 \times 10^6}{1000} = 4500$ pet unicorns in Ireland.