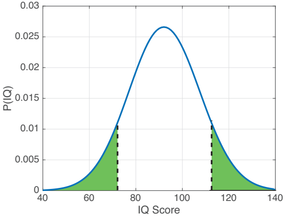
Overview

- Confidence Intervals
- Sampling and Opinion Polls
- Error Correcting Codes
- Number of Pet Unicorns in Ireland

Confidence Intervals

- When a random variable lies in an interval $a \leq X \leq b$ with a specified probability we call this a **confidence interval** e.g. $p - 0.05 \leq Y \leq p + 0.05$ with probability at least 0.95.



Confidence Intervals: Using Inequalities

- Chebyshev inequality allows us to calulate confidence intervals given the mean and variance of a random variable.
- For sample mean $\bar{X} = \frac{1}{N}\sum_{k=1}^{N} X_k$, Chebyshev inequality tells us $P(|\bar{X} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{N\epsilon^2}$ where μ is mean of X_k and σ^2 is its variance.
- E.g. When $\epsilon = \frac{\sigma}{\sqrt{0.05N}}$ then $\frac{\sigma^2}{N\epsilon^2} = 0.05$ and Chebyshev tells us that $\mu - \frac{\sigma}{\sqrt{0.05N}} \leq \bar{X} \leq \mu + \frac{\sigma}{\sqrt{0.05N}}$ with probability at least 0.95.

Confidence Intervals: Using CLT

- CLT also allows us to calculate an approximate confidence interval. When $X \sim N(\mu, \sigma^2)$:

$$P(-\sigma \leq X - \mu \leq \sigma) \approx 0.68$$
$$P(-2\sigma \leq X - \mu \leq 2\sigma) \approx 0.95$$
$$P(-3\sigma \leq X - \mu \leq 3\sigma) \approx 0.997$$

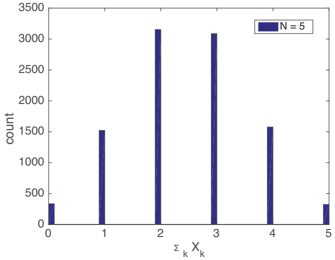
- These are 1σ, 2σ, 3σ confidence intervals
- $\mu \pm 2\sigma$ is the 95 confidence interval for a Normal random variable with mean μ and variance σ^2. In practice often use either $\mu \pm \sigma$ or $\mu \pm 3\sigma$ as confidence intervals.
- Recall claim by Goldman Sachs that crash was a 25σ event (expected to occur once in 10^{135} years[1]) ?

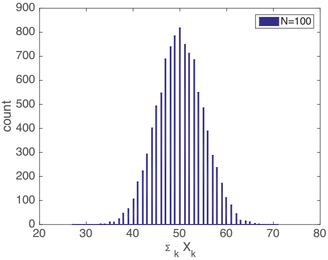
[1] http://arxiv.org/pdf/1103.5672.pdf

Confidence Intervals

But ...

- This confidence interval differs from from that derived from Chebyshev inequality. Chebyshev confidence intervals are actual confidence intervals. Those derived from CLT are only approximate (accuracy depends on how large N is)
- We need to be careful to check that N is large enough that distribution really is almost Gaussian. This might need large N.
- Recall coin toss example:





Example: Running Time of New Algorithm

Suppose we have an algorithm to test. We run it N times and measure the time to complete, gives measurements $X_1, \cdots, X_N$.

- Mean running time is $\mu = 1$, variance is $\sigma^2 = 4$
- How many trials do we need to make so that the measured sample mean running time is within 0.5s of the mean μ with 95% probability ? $P(|X - \mu| \geq 0.5) \leq 0.05$ where $X = \frac{1}{N} \sum_{k=1}^{N} X_k$
- CLT tells us that $X \sim N(\mu, \frac{\sigma^2}{N})$ for large N. Normal distribution satisfies the "68-95-99.7 rule".

$$P(-\sigma \leq X - \mu \leq \sigma) \approx 0.68$$
$$P(-2\sigma \leq X - \mu \leq 2\sigma) \approx 0.95$$
$$P(-3\sigma \leq X - \mu \leq 3\sigma) \approx 0.997$$

So we need $2\sigma = 2\sqrt{\frac{\sigma^2}{N}} = 0.5$ i.e. $N \geq 64$.

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.
- 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 $X = nY$.
- How do we select n so that estimate is not more than 5% different the mean 95% of the time ? That is, $P(|Y - p| \geq 0.05) \leq 0.05$

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

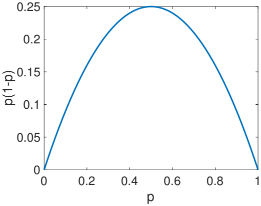
Fri, Feb 26, 2016, 22:48 | Updated: Sat, Feb 27, 2016, 08:05

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.
- Use the CLT to estimate confidence interval:
  - X_i is Bernoulli random variable, so mean p, variance $p(1-p)$. Variance of Y is $\sigma^2 = p(1-p)/n$ with $n = 5260$.
  - We don't know p, but let's plot $p(1-p)$ vs p:



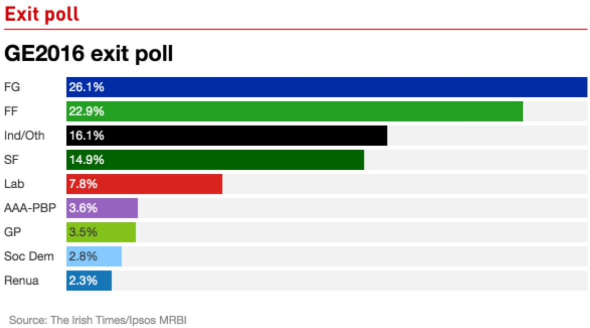
  - Assume Y is normally distributed, then

$$P(-\sigma \leq Y - p \leq \sigma) \approx 0.68$$
$$P(-2\sigma \leq Y - p \leq 2\sigma) \approx 0.95$$

  - Use $\sigma^2 \leq 0.25/5260$, i.e. $\sigma \leq 0.0069$. Then $2\sigma \leq 0.0138$ or 1.38%

Sampling: Irish Election 2016

How did exit poll predictions compare with final results ?



Seats by party

Party	FG	FF	SF	IO	LAB	AAA-PBP	SD	GP	RI	TBC
Seats	49	44	22	21	6	5	3	2	0	6
% 1st Pref	25.52%	24.35%	13.85%	17.83%	6.61%	3.95%	3.00%	2.72%	2.18%	--

152 of 158 seats filled **37 of 40** constituencies complete **65%** national turnout

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 of 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

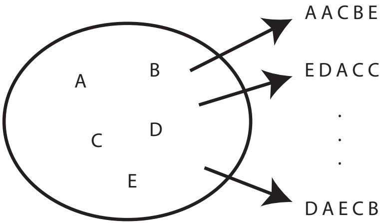
[2] E.g. https://en.wikipedia.org/wiki/Shy_Tory_Factor and https://en.wikipedia.org/wiki/Bradley_effect

Confidence Intervals: Bootstrapping

Mean μ, variance σ^2 and number of points N summarises our N data points using three numbers. But we have $N \gg 3$ data points. Can we use these to also empirically estimate the distribution of the sample mean ? Yes !

- Makes use of the fact that computing power is cheap.
- The N data points are drawn independently from the same probability distribution F
- So the idea is to use these N data points as a surrogate for F. To generate new samples from F we draw uniformly at random from our N data points. This is **sampling with replacement**.
- Suppose our data is $\{A, B, C, D, E, F\}$. Select one point uniformly at random e.g. B. Select a second point uniformly at random, might be B again or might be something else. And so on until we get desired number of samples.
- Bootstrapping is an example of a **resampling method**

Confidence Intervals: Bootstrapping



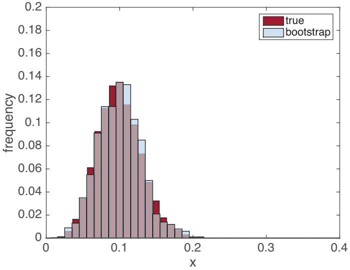
Bootstrapping:

- Draw a sample of N data points uniformly at random from data, with replacement.
- Using this sample estimate the mean $X_1 = \frac{1}{N} \sum_{i=1}^{N} X_{1,i}$
- Repeat, to generate a set of estimates X_1, $X_2, \cdots$.
- The distribution of these estimates approximates the distribution of the sample mean (its not exact)

Confidence Intervals: Bootstrapping

Example: coin tosses yet again.

- Toss $N = 100$ biased coins, lands heads with prob $p = 0.1$. $X_i = 1$ if i'th toss is heads, 0 otherwise
- Sample with replacement from the $N = 100$ data points.
- Calculate sample mean $X = \frac{1}{N}\sum_{i=1}^{N} X_i$
- Repeat 1000 times and plot observed distribution of X.



Confidence Intervals: Bootstrapping

Matlab code used to generate above plot:

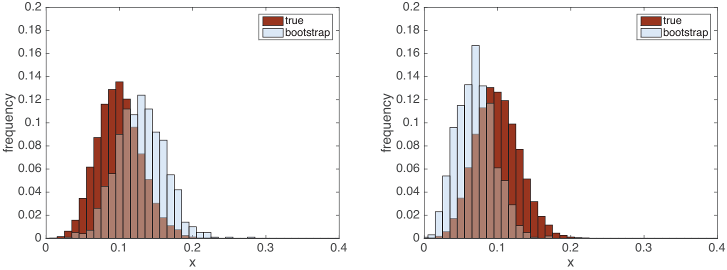
```
p=0.1; n=100;
y=[];
for i=1:10000,
    r=rand(1,n); x=(r<=p); y=[y;sum(x)/n];
end;
xx=[0:1/n-eps:1]; nn=histc(y,xx,1);
hold off, bar(xx,nn/sum(nn),1)

r=rand(1,n); x=(r<=p);
y1=[];
for i=1:1000,
    y1= [y1; sum(x(randi(n,1,n)))/n];
end
nn1=histc(y1,xx);
hold on, bar(xx,nn1/sum(nn1),1,'g')
axis([0 0.4 0 0.2])
```

Confidence Intervals: Bootstrapping

Note: bootstrap estimate of the distribution is only approximate.

- Different data leads to different estimates of the the distribution, e.g. here are two more runs of the coin toss example.
- But v handy all the same.
- Using our empirical estimate of the distribution of the sample mean X we can estimate confidence intervals etc.



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.

Number of Pet Unicorns in Ireland: Confidence Intervals

- Data consists $N = 1000$ survey results. $X_i = 1$ if answer "yes" and 0 otherwise. We have 999 of the X_i's equal to 0 and one equal to 1.
- Sample mean is $\frac{1}{1000} = 1 \times 10^{-3}$, variance is $\frac{1}{1000} - (\frac{1}{1000})^2 = 9.99 \times 10^{-4}$.
- Normal approximation suggests:

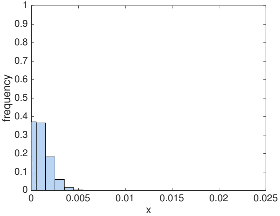
$$P(-2\sigma \leq X - \mu \leq 2\sigma) \approx 0.95$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$$
$$P(-9.98 \times 10^{-4} \leq X \leq 3 \times 10^{-3}) \approx 0.95$$

- Scaling by 4.5M (approx population of Ireland), we estimate number of per unicorns lies in the range -4500 to $13,500$ with 95% probability. This interval includes zero, so we're not too confident that there are in fact any pet unicorns.

Confidence Intervals: Confidence Intervals

Bootstrapping doesn't require data to be normally distributed.

- Data consists $N = 1000$ survey results. $X_i = 1$ if answer "yes" and 0 otherwise. We have 999 of the X_i's equal to 0 and one equal to 1. Bootstrap estimate of distribution of $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$:



- We can see that number of unicorns is likely to be 0.
- Estimate $P(\bar{X} < 1/4.5 \times 10^3) \approx 0.4$ and $P(\bar{X} < 10/4.5 \times 10^3) \approx 0.9$ i.e. if scale up up $4.5 \times 10^6/N = 4.5 \times 10^3$ then prob no pet unicorns is estimated at about 0.4 and of less than 10 unicorns at about 0.9.
- Compare with original estimate of 4500 pet unicorns based on mean value alone.

Confidence Interval Wrap-up

We have three different approaches for estimating confidence intervals: (i) Chebyshev (and other(Inequalities, (ii) Bootstrapping and (iii) CLT. Each has pros and cons:

- CLT: $\bar{X} \sim N(\mu, \frac{\sigma^2}{N})$ as $N \to \infty$
  - Gives full distribution of $\bar{X}$
  - Only requires mean and variance to fully describe this distribution
  - But is an approximation when N finite, and hard to be sure how accurate (how big should N be ?)
- Chebyshev (and other) inequalities:
  - Provide an actual bound (not an approximation)
  - Works for all N
  - But loose in general.
- Bootstrapping:
  - Gives full distribution of $\bar{X}$, doesn't assume Normality.
  - But is an approximation when N finite, and hard to be sure how accurate (how big should N be ?)
  - Requires availability of all N measurements.