Overview

- Distribution of Sample Mean
- Weak Law of Large Numbers
- Central Limit Theorem
- Confidence Intervals (Again)
- Confidence Intervals: Boostrapping
Consider \( N \) random variables \( X_1, \cdots, X_N \).

- Lets revisit our old friend \( \bar{X} = \frac{1}{N} \sum_{k=1}^{N} X_k \).
- \( \bar{X} \) is called the “sample mean” or the “empirical mean”.
- \( \bar{X} \) is a random variable.

Suppose we observe values for \( X_1, \cdots, X_N \) and calculate the empirical mean of the observed values. That gives is one value for \( \bar{X} \). But the value of \( \bar{X} \) changes depending on the observed values.

- Suppose we toss a fair coin \( N = 5 \) times and get \( H, H, H, T, T \). Let \( X_k = 1 \) when come up heads. Then \( \frac{1}{N} \sum_{k=1}^{N} X_k = \frac{3}{5} \)
- Suppose we toss the coin another \( N = 5 \) times and get \( T, T, H, T, H \). Now \( \frac{1}{N} \sum_{k=1}^{N} X_k = \frac{2}{5} \)
Distribution of Sample Mean

Toss fair coin $N = 5$ times and calculate $\sum_{k=1}^{N} X_k$. Repeat, and plot histogram of values. Its binomial $\text{Bin}(N, \frac{1}{2})$.

- $X=[]; \text{for } i=1:10000, X=[X, \text{sum}((\text{rand}(1,5)<0.5))] \text{; end; hist}(X,50)$
Random variable $\bar{X} = \frac{1}{N} \sum_{k=1}^{N} X_k$.

- Suppose the $X_k$ are independent and identically distributed
- Each $X_k$ has mean $E(X_k) = \mu$ and variance $\text{var}(X_k) = \sigma^2$.

Then we can calculate the mean of $\bar{X}$ as:

$$E(\bar{X}) = E\left(\frac{1}{N} \sum_{k=1}^{N} X_k\right) = \frac{1}{N} \sum_{k=1}^{N} E(X_k) = \mu = \frac{1}{N^2} \sum_{k=1}^{N} \text{var}(X_k) = \frac{N \sigma^2}{N^2} = \frac{\sigma^2}{N}$$

- We say $\bar{X}$ is an **unbiased estimator** of $\mu$ since $E[\bar{X}] = \mu$
Distribution of Sample Mean

We can calculate the variance of $\bar{X}$ as:

$$
\text{var}(\bar{X}) = \text{var}\left(\frac{1}{N} \sum_{k=1}^{N} X_k\right) = \frac{1}{N^2} \text{var}\left(\sum_{k=1}^{N} X_k\right)
$$

$$
= \frac{1}{N^2} \sum_{k=1}^{N} \text{var}(X_k) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}
$$

- As $N$ increases, the variance of $\bar{X}$ falls.
- Recall that $\text{Var}(N X) = N^2 \text{Var}(X)$ for random variable $X$.
- But when add together independent random variables $X_1 + X_2 + \cdots$ the variance is only $N \text{Var}(X)$ rather than $N^2 \text{Var}(X)$.
- This is due to statistical multiplexing. Small and large values of $X_i$ tend to cancel out for large $N$. 
Weak Law of Large Numbers

Consider $N$ independent identically distributed random variables $X_1, \cdots, X_N$ each with mean $\mu$ and variance $\sigma^2$. Let $\bar{X} = \frac{1}{N} \sum_{k=1}^{N} X_k$. For any $\epsilon > 0$:

$$P(|\bar{X} - \mu| \geq \epsilon) \to 0 \text{ as } N \to \infty$$

That is, $\bar{X}$ concentrates around the mean $\mu$ as $N$ increases.

Proof:

- $E(\bar{X}) = E(\frac{1}{N} \sum_{k=1}^{N} X_k) = \frac{1}{N} \sum_{k=1}^{N} E(X_k) = \mu$
- $\text{var}(\bar{X}) = \text{var}(\frac{1}{N} \sum_{k=1}^{N} X_k) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$
- By Chebyshev’s inequality: $P(|\bar{X} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{N\epsilon^2}$
Who cares?

- Suppose we have an event $E$
- Define random variable $X_i$ equal to 1 when event $E$ is observed in trial $i$ and 0 otherwise
- Recall $E[X_i] = P(E)$ is the probability that event $E$ occurs.
- $\bar{X} = \frac{1}{N} \sum_{k=1}^{N} X_k$ is then the relative frequency with which event $E$ occurs
- And ... $P(|\bar{X} - \mu| \geq \epsilon) \rightarrow 0$ as $N \rightarrow \infty$
tells us that this relative frequency $\bar{X}$ converges to the probability $P(E)$ of event $E$
- So its the law of large numbers that formalises the intuition of probability as frequency when an experiment can be repeated many times.
Central Limit Theorem (CLT)

But can we say more about the probability distribution of $\bar{X}$?

- Coin toss example again, but now we plot a histogram of $\bar{X}$ as $N$ increases.

- See that (i) curve narrows as $N$ increases, it “concentrates”.
- Curve is roughly “bell-shaped”
Central Limit Theorem (CLT)

Matlab code used to generate above plot:

```matlab
figure(1), clf, hold on
for N=[500,5000,20000],
    X=[];
    for i=1:10000,
        X=[X, sum((rand(1,N)<0.5))/N];
    end;
    [n,x]=hist(X,100); n=n/trapz(x,n);
    stairs(x,n)
end
axis([0.4 0.6 0 140])
xlabel('(Σ_{k=1}^{N} X_k)/N'), ylabel('frequency')
legend('N=500', 'N=5000', 'N=20000')
```
Central Limit Theorem (CLT)

Consider $N$ independent identically distributed random variables $X_1, \cdots, X_N$ each with mean $\mu$ and variance $\sigma^2$. Let $\bar{X} = \frac{1}{N} \sum_{k=1}^{N} X_k$. Then:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{N}) \text{ as } N \to \infty$$

- This says that as $N$ increases the distribution of $\bar{X}$ converges to a Normal (or Gaussian) distribution. This is the bell shape that we saw.
- The distribution has mean $\mu$ and variance $\sigma^2/N$.
- Variance $\sigma^2/N \to 0$ as $N \to \infty$. So distribution concentrates around the mean $\mu$ as $N$ increases.
CLT vs Chernoff

- CLT: \( \tilde{X} \sim N(\mu, \frac{\sigma^2}{N}) \) as \( N \to \infty \)
  - Gives full distribution of \( \tilde{X} \)
  - Only requires mean and variance to fully describe this distribution
  - But is an approximation when \( N \) finite, and hard to be sure how accurate
- Chernoff: \( P(\tilde{X} \geq a) \leq e^{-ta}e^{\log E[e^{t\tilde{X}}]}, \ t > 0 \)
  - Its an actual bound (not an approximation)
  - Works for all \( N \)
  - But loose in general.
- Both have their uses ...
Central Limit Theorem (CLT)

Big advantage of CLT is that it is tight (provided \( N \) is large enough). Convergence is pointwise onto normal distribution.
Distribution of Sample Mean

Matlab code used to generate above plot:

```matlab
figure(1), clf, hold on
for N=[500,5000,20000],
    X=[];
    for i=1:10000,
        X=[X,sum((rand(1,N)<0.5))/N];
    end;
    [n,x]=hist(X,100); n=n/trapz(x,n);
    z=[0:0.005:1]; sigma=0.25/N;
    clt=exp(-(z-0.5).^2/(2*sigma));
    clt=clt/(sqrt(2*pi*sigma));
    plot(x,n,z,clt,'--')
end
axis([0.4 0.6 0 140])
xlabel('(\Sigma_k X_k)/N'), ylabel('frequency')
legend('Sample mean','N(0.5,0.25/N)')
```
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 \( \mu \) 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. \( N \geq 64 \).
Confidence Intervals (Again)

• Recall that when a random variable lies in an interval $a \leq X \leq b$ with a specified probability we call this a 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 $1\sigma$, $2\sigma$, $3\sigma$ confidence intervals

• $\mu \pm 2\sigma$ is the 95 confidence interval for a Normal random variable with mean $\mu$ and variance $\sigma^2$. 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 $25\sigma$ event (expected to occur once in $10^{135}$ years$^1$)?

Confidence Intervals (Again)

But ...

- These confidence intervals differ from those we previously derived from Chernoff inequality. Chernoff 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:

- We also need that the $X_i$ are independent (similarly with Chernoff and pals). Likely violated by financial markets?
Confidence Intervals: Bootstrapping

Sample mean summarises our \( N \) data points in a single number. But we have \( N \) 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 (it’s 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 $X$. 

![Histogram of true and bootstrap distribution of X]
Confidence Intervals: Bootstrapping

Matlab code used to generate above plot:

```matlab
1 p = 0.1; n = 100;
2 y = []; 
3 for i = 1:10000,
4     r = rand(1,n); x = (r <= p); y = [y; sum(x)/n];
5 end;
6 xx = [0:1/n-eps:1]; nn = histc(y, xx, 1);
7 hold off, bar(xx, nn/sum(nn), 1)
8
9 r = rand(1,n); x = (r <= p);
10 y1 = []; 
11 for i = 1:1000,
12     y1 = [y1; sum(x(randi(n,1,n)))/n];
13 end
14 nn1 = histc(y1, xx);
15 hold on, bar(xx, nn1/sum(nn1), 1, 'g')
16 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.
Confidence Intervals: Bootstrapping

Bootstrapping doesn’t require data to be normally distributed.

- How about those pet unicorns?
- 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. This is very non-normal.
- Sample mean is $\frac{1}{1000}$. But bootstrap estimate of the distribution of this value makes clear that its at least as likely to be 0.

- Beware strange predictions e.g. cost of cybercrime $>400B$. What do you reckon?