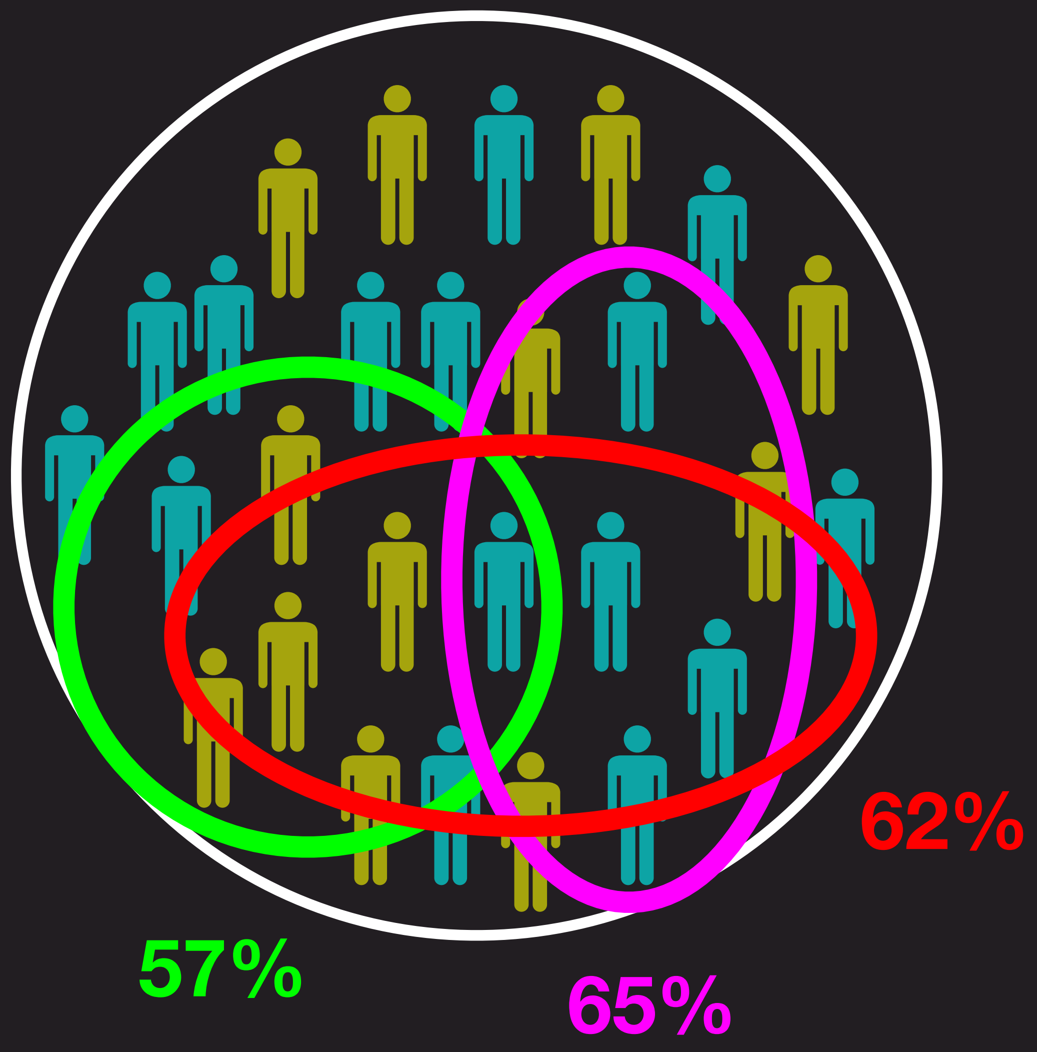
Opinion poll



Candidate A has 60% support

Candidate B has 40% support

We do not know these values

How do we determine the true numbers?

Is it practical to ask EVERY person whom they support?

We sample a few people

How close are these numbers to the real value of 60%?

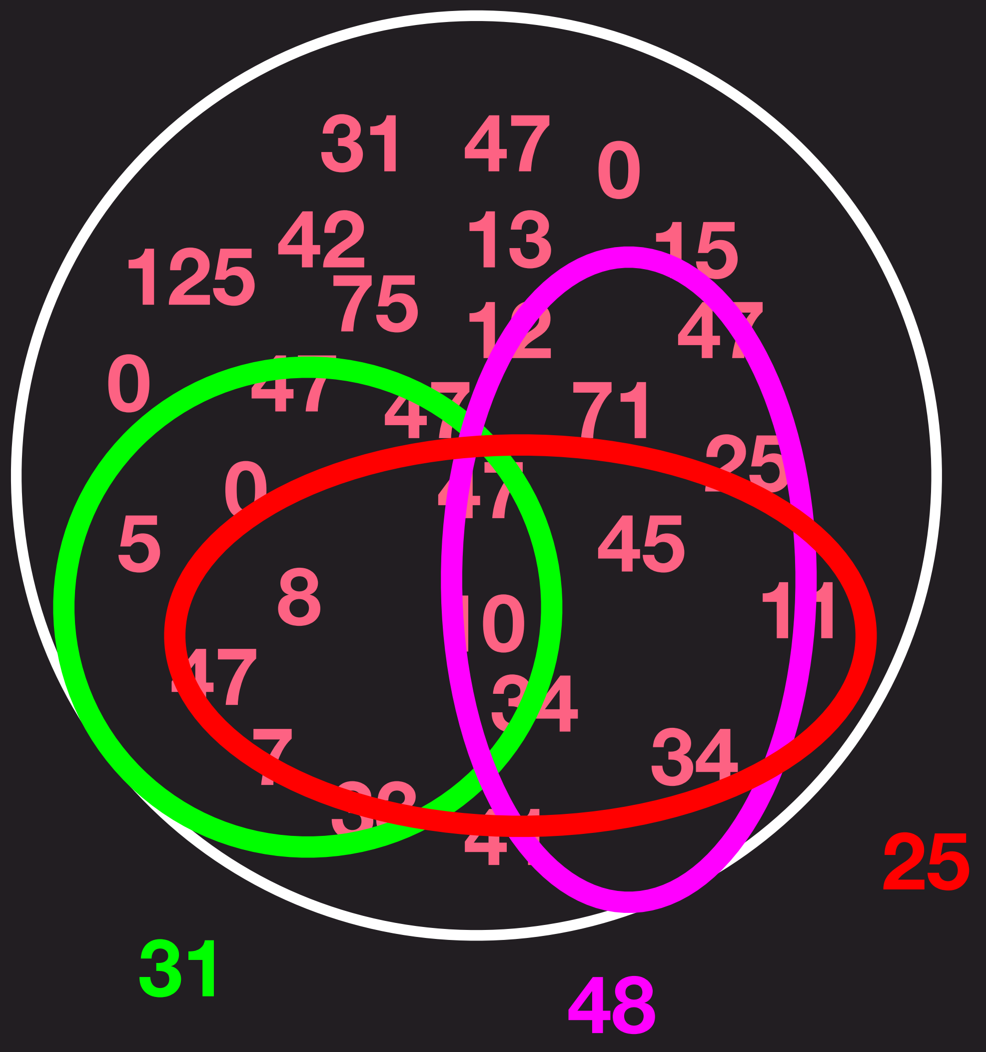
This depends on the number of people we have asked

This number we will call “n” - the number of samples

It is true than as “n” increases, the accuracy increases

But budget constraints put an upper limit on “n”

Sehwag's Runs



Suppose we watch 10 or 20 matches and guess his average

How close is the "sample mean" to the true mean/average

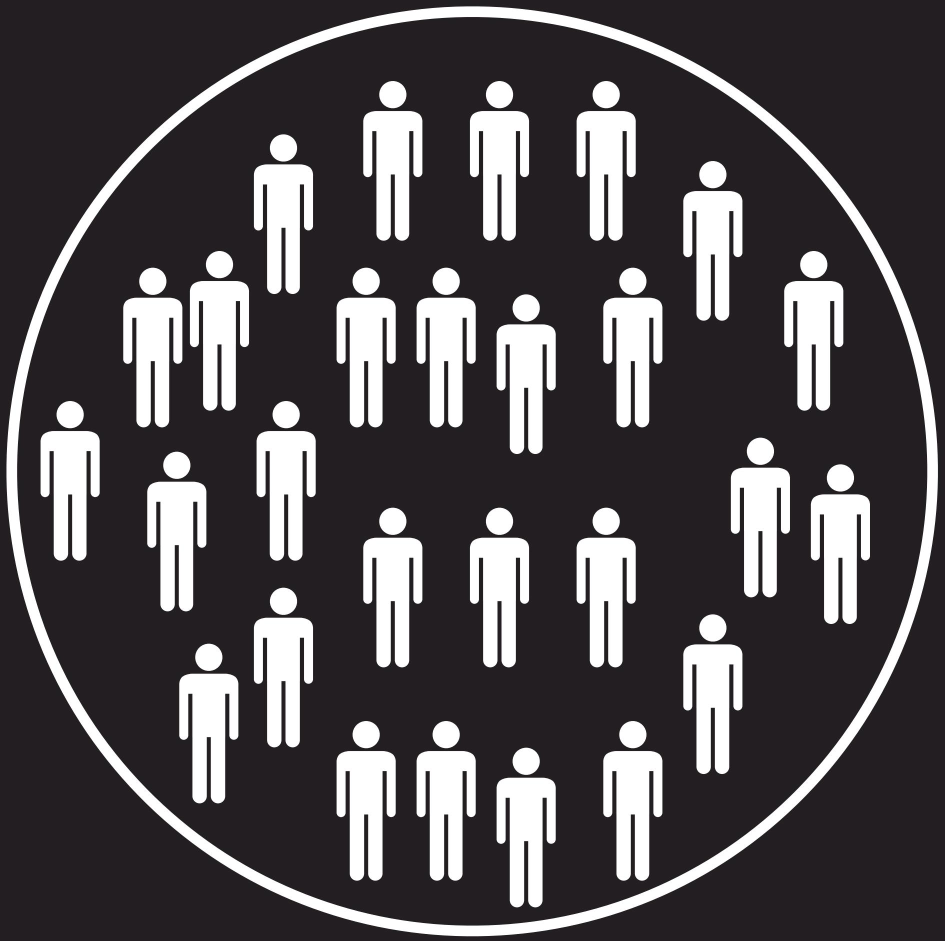
To answer this, we need to know some details of the sample mean

These numbers (31, 48, 25 etc) are sample means

These numbers have their own mean, variance, histogram, etc

We need to make statistically relevant remarks on the true mean using these sample means

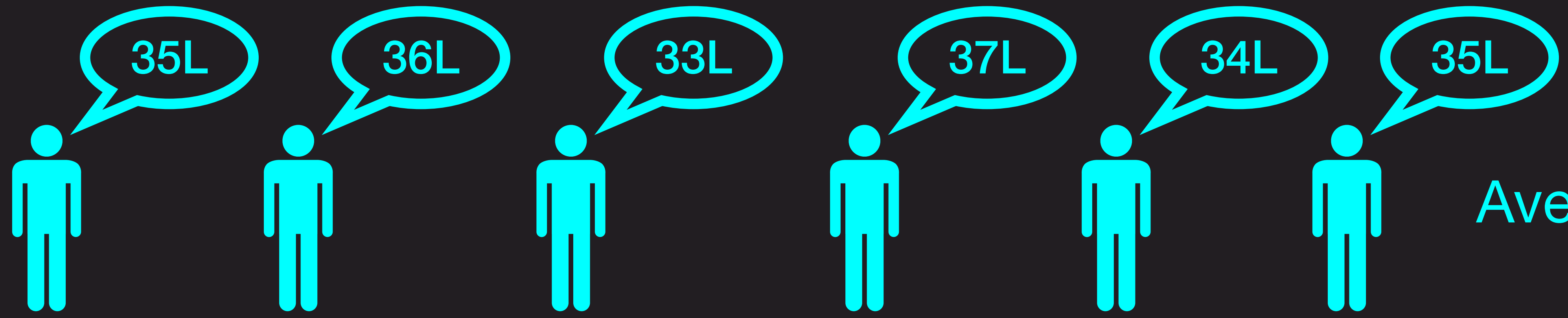
Confidence Intervals



SDE-2 Salary

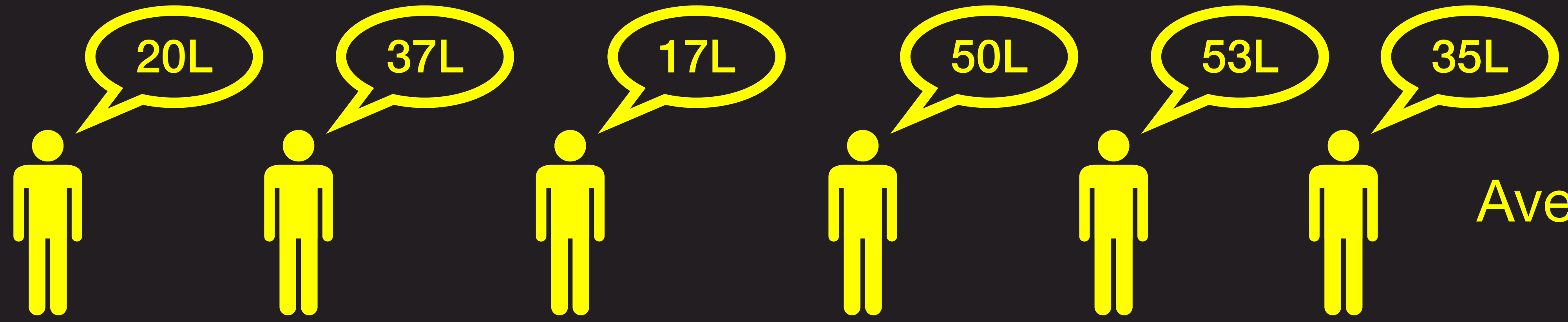
You want to know what is the average salary of all SDE-2

Survey 1 Results of a small survey is here



Average = 35 L

Survey 2 Results of another small survey has also come



Average = 35 L

Both surveys have the same mean/average

In which are you more confident? Survey 1

Let us quantify this confidence

Confidence Intervals

SDE-2 Salary

Survey 1

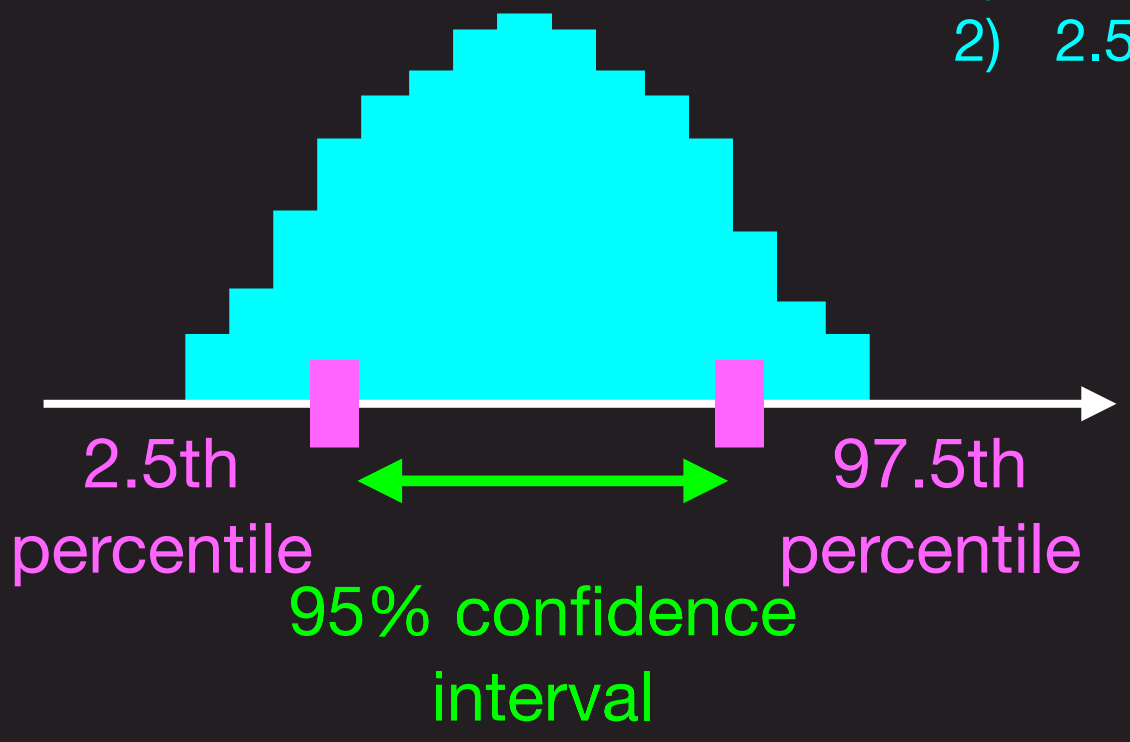
[35, 36, 33, 37, 34, 35]

Bootstrapped samples	Bootstrapped mean
[33, 35, 37, 33, 34, 35]	34.5
[36, 36, 37, 35, 34, 35]	35.5
[35, 35, 35, 35, 35, 34]	34.83
[34, 37, 33, 36, 35, 37]	35.33
[35, 35, 35, 33, 35, 33]	34.33

10000 times

To get the 95% confidence interval, we need

1) 97.5th percentile
2) 2.5th percentile

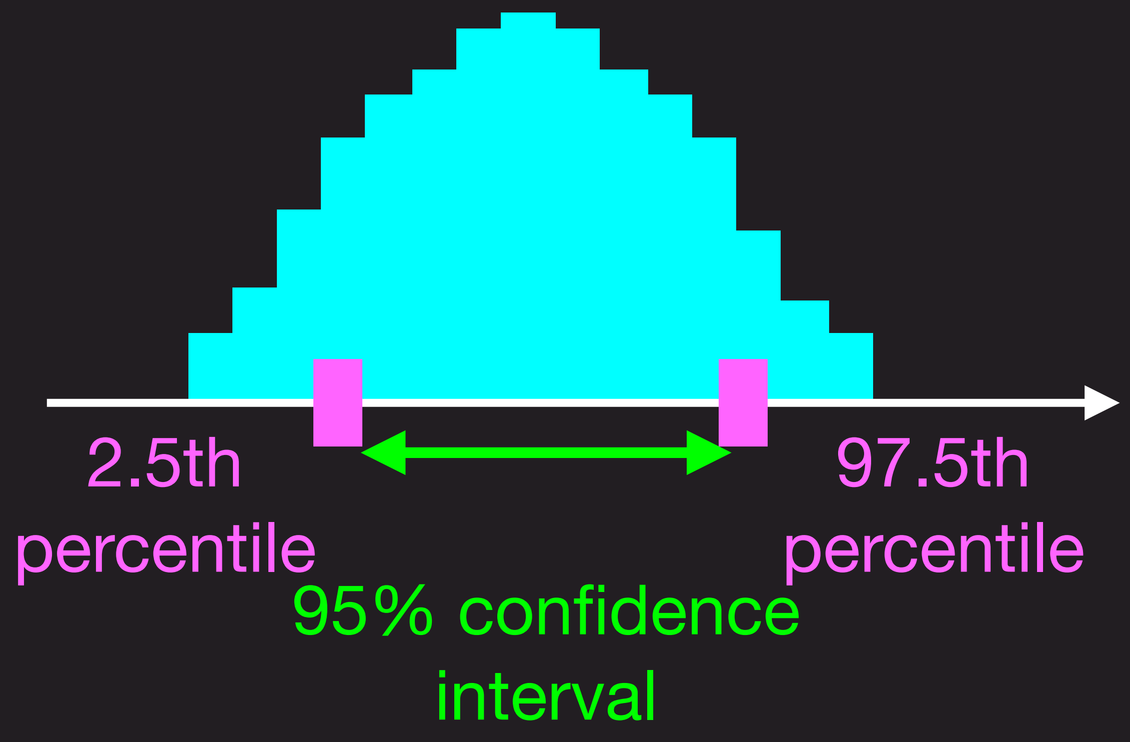


Confidence Intervals

SDE-2 Salary

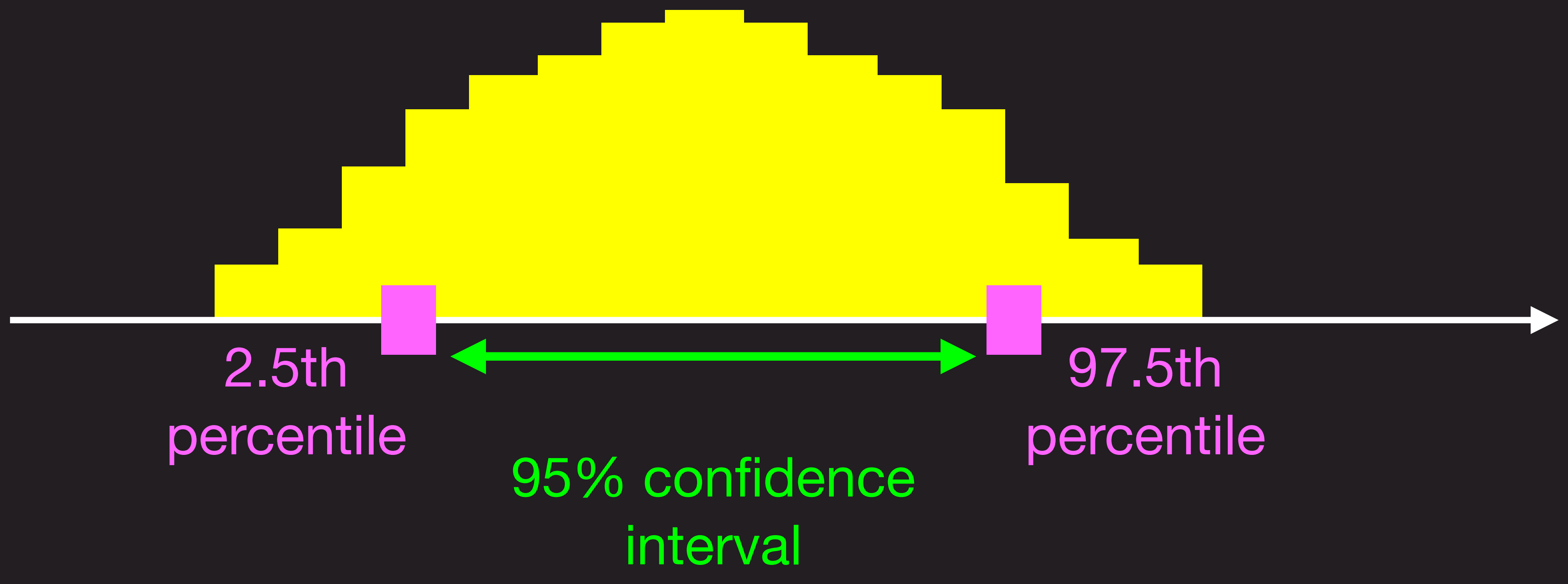
Survey 1

[35, 36, 33, 37, 34, 35]

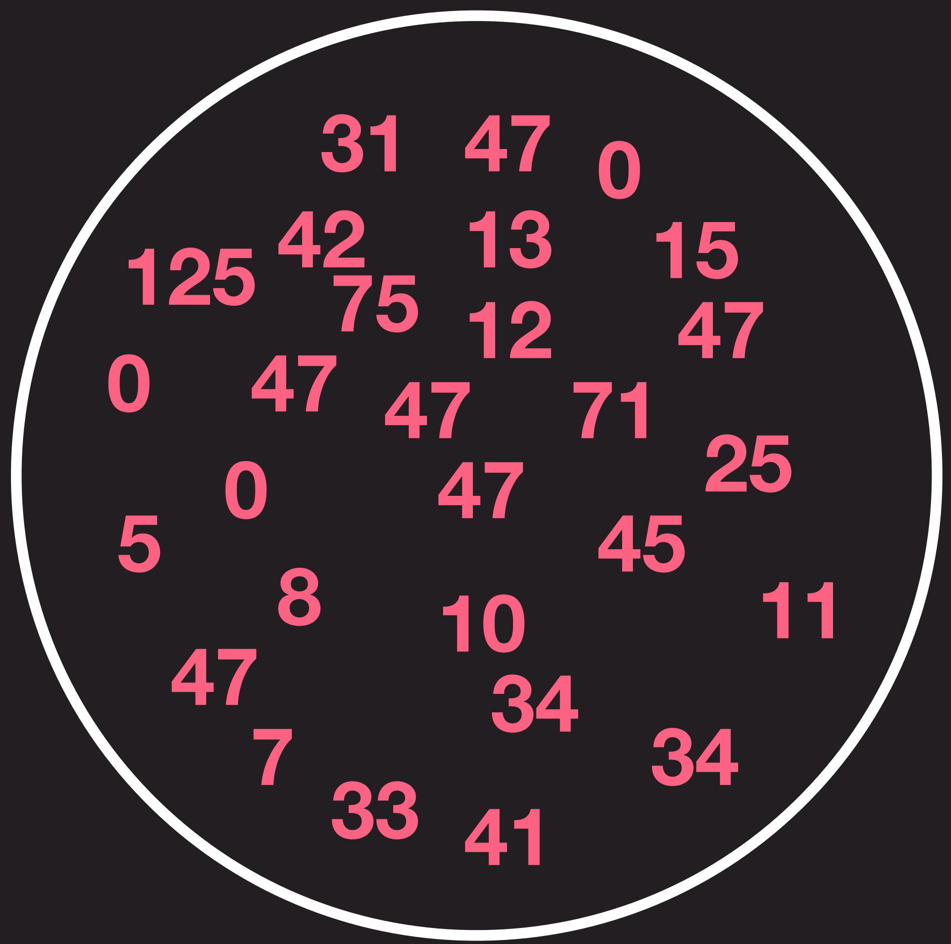


Survey 2

[20, 37, 17, 50, 53, 33]



Sehwag's Runs



Sample means

31	m_1
48	m_2
25	m_3

μ True mean of all the matches

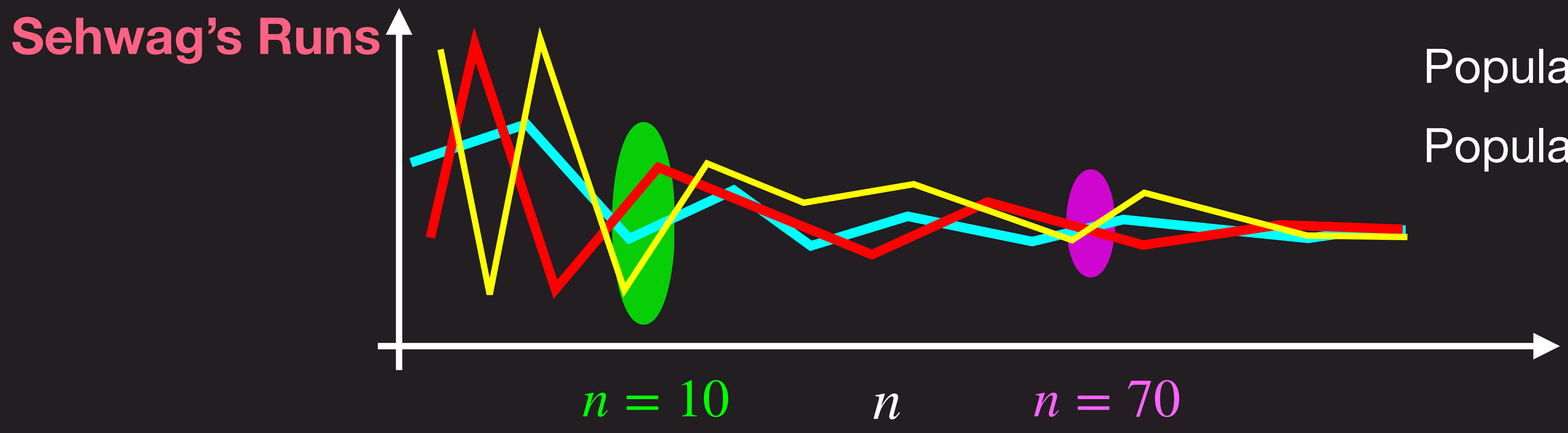
"Population mean"

33.77

σ True standard deviation of all the matches

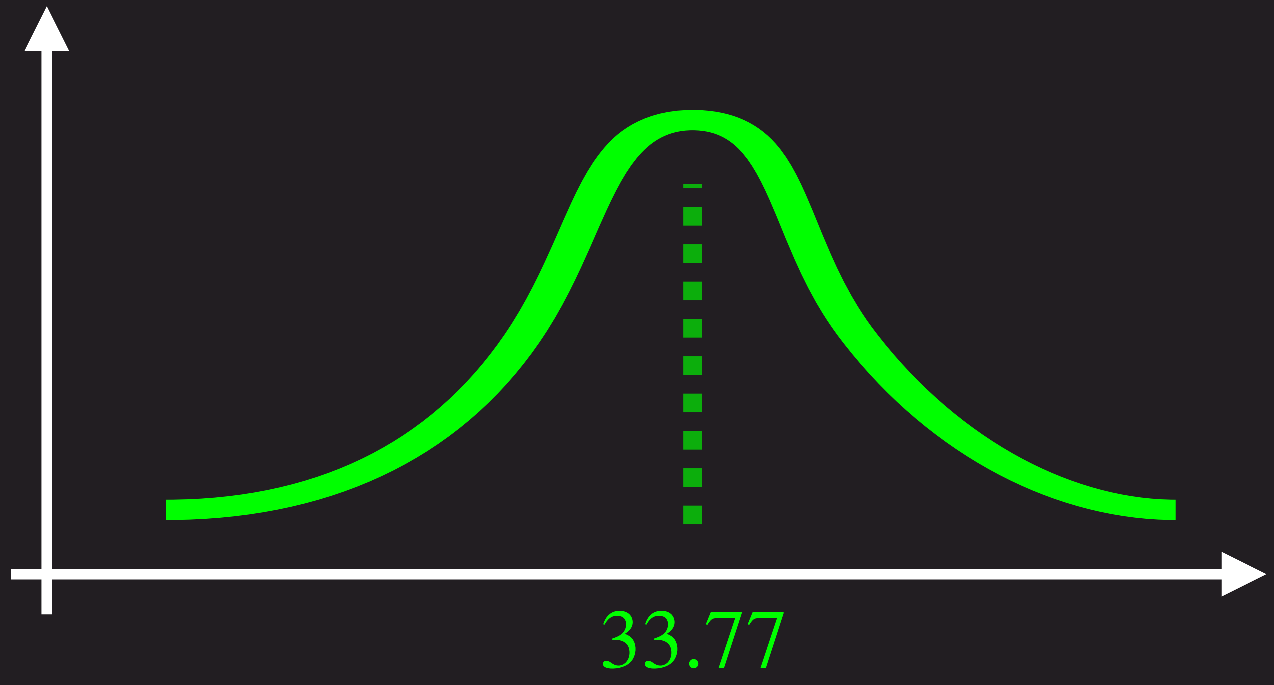
"Population standard deviation"

34.81

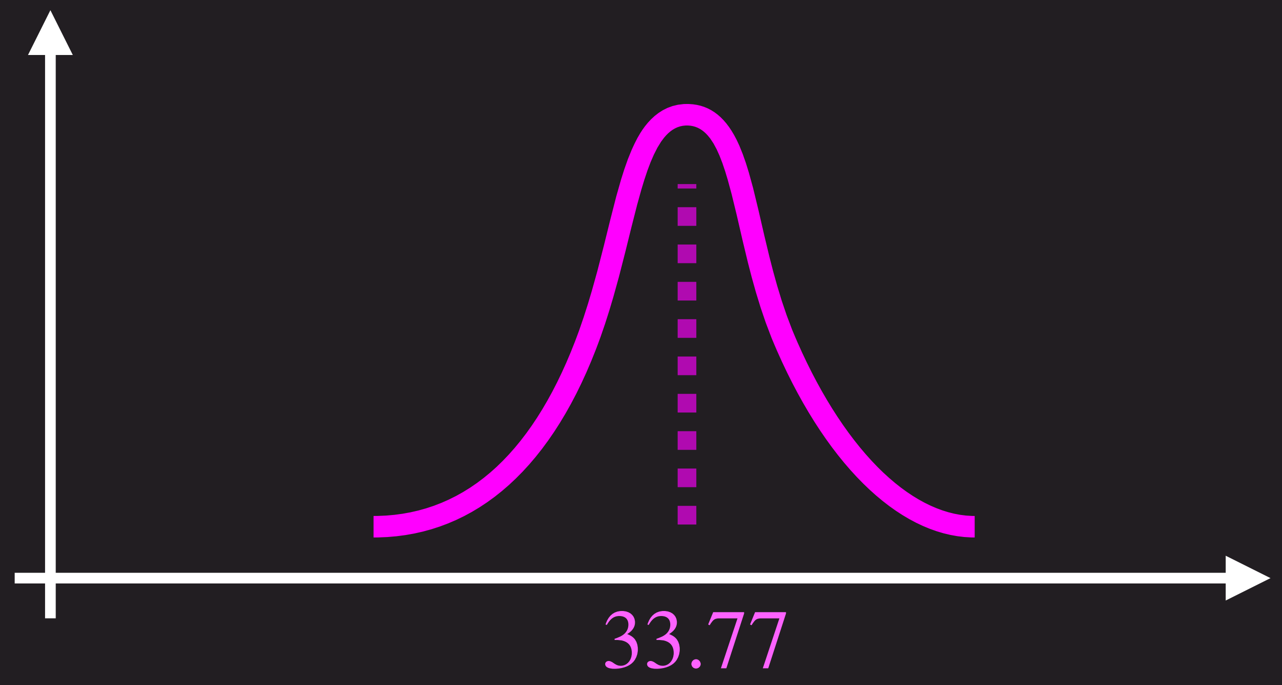


Population mean $\mu = 33.77$

Population std dev $\sigma = 34.81$

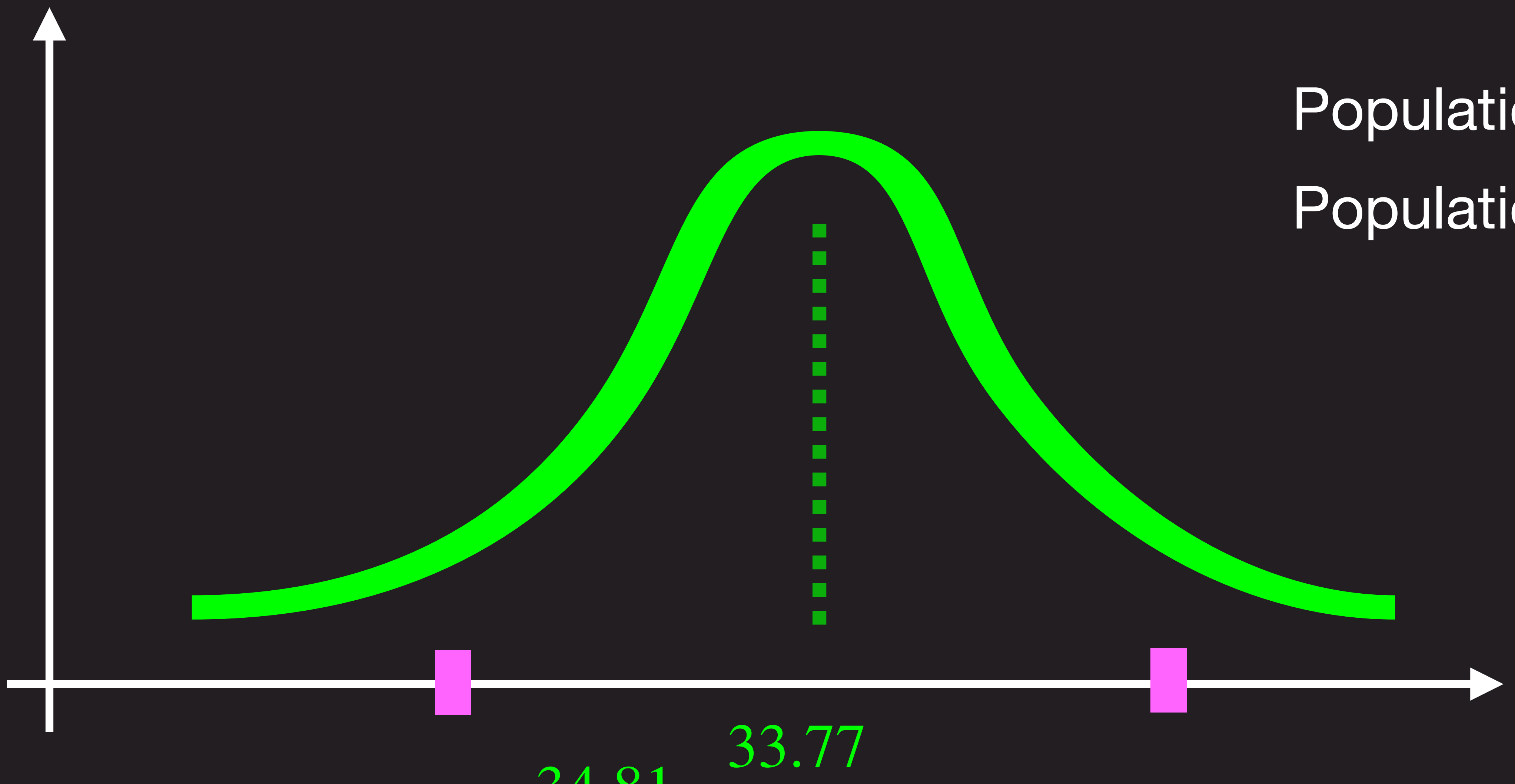


Std dev = $\dfrac{34.81}{\sqrt{10}}$



Std dev = $\dfrac{34.81}{\sqrt{70}}$

Sehwag's Runs



Population mean $\mu = 33.77$

Population std dev $\sigma = 34.81$

Std dev = $\dfrac{34.81}{\sqrt{n}}$

"Standard Error"

To compute the 95% confidence interval, we need Z-score of 0.975 and 0.025

`norm.ppf(0.025)` $= -1.96$

`norm.ppf(0.975)` $= 1.96$

If the sample mean of "n" samples is, for example, 32, then we say

Confidence interval $= \left[32 - \dfrac{1.96 * 34.81}{\sqrt{n}},\ 32 + \dfrac{1.96 * 34.81}{\sqrt{n}}\right]$

Central Limit Theorem

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

(Sample mean)

(Eg: X_i Schwag scores
μ: population mean (33.7)
σ: pop. std dev (34.8)

$\bar{X}$ has a Gaussian distribution

mean of $\bar{X}$ $\quad E[\bar{X}] = \mu$ $\rightarrow$ same as pop. mean

Std dev of $\bar{X}$ $= \dfrac{\sigma}{\sqrt{n}}$