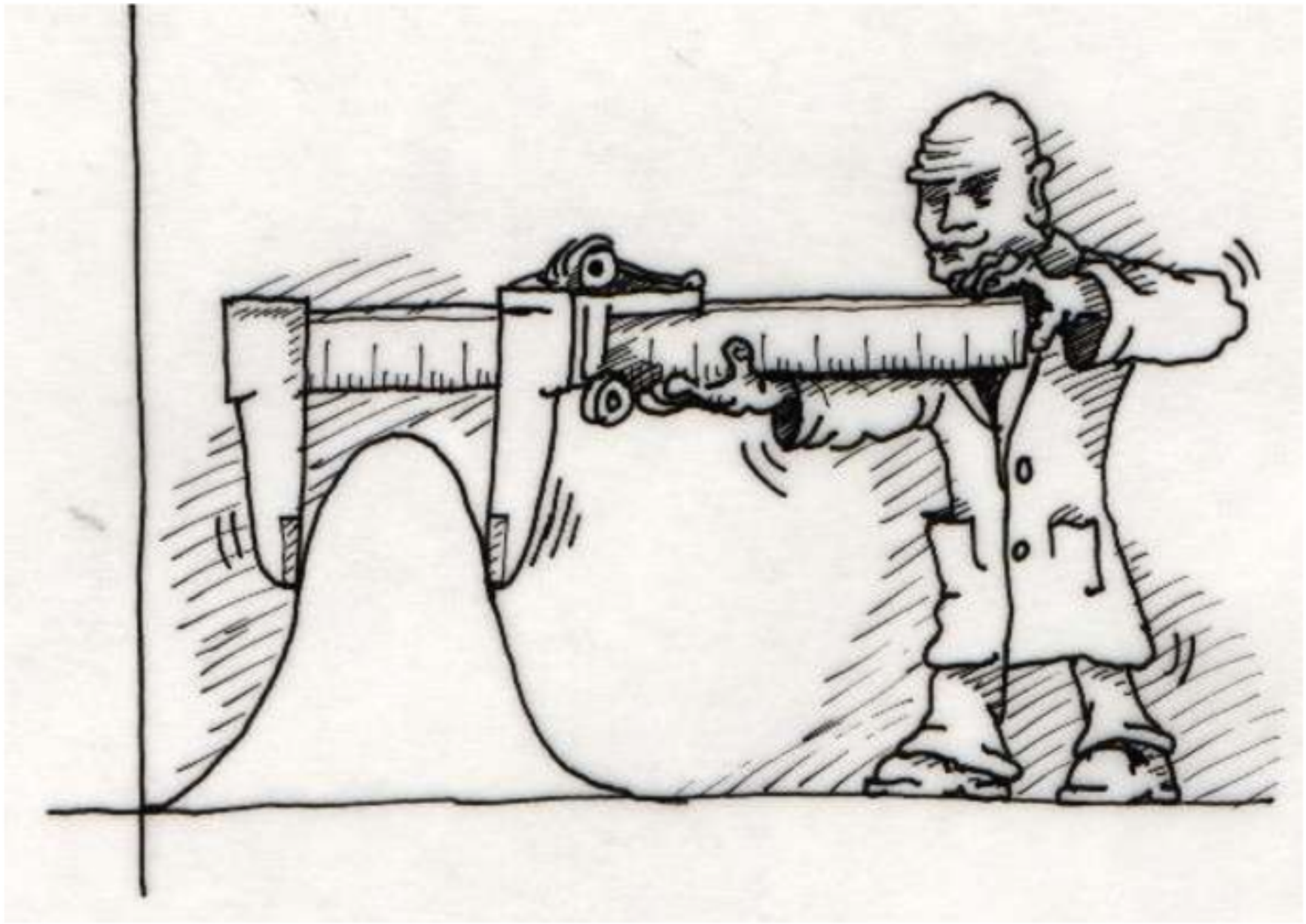
ALL YOU NEED TO KNOW ABOUT STATISTICS

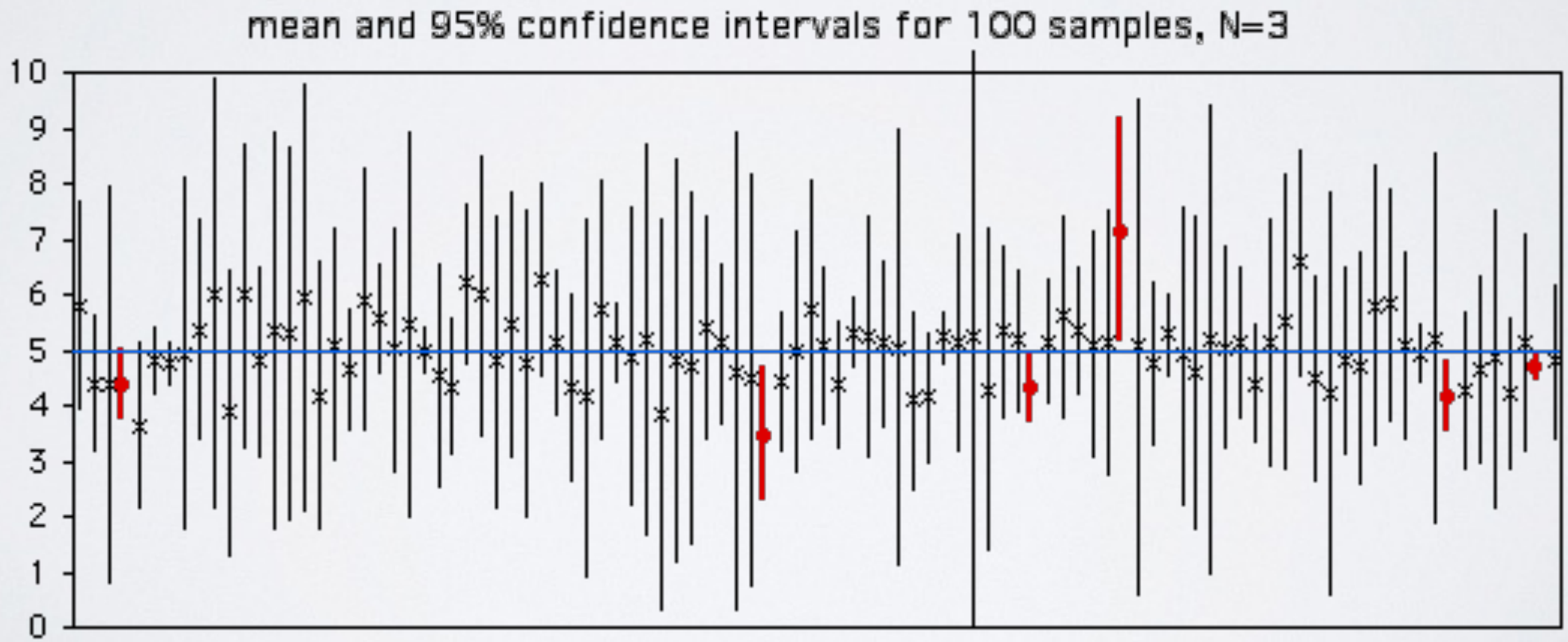
In 15 minutes

Roberto A. Vitillo

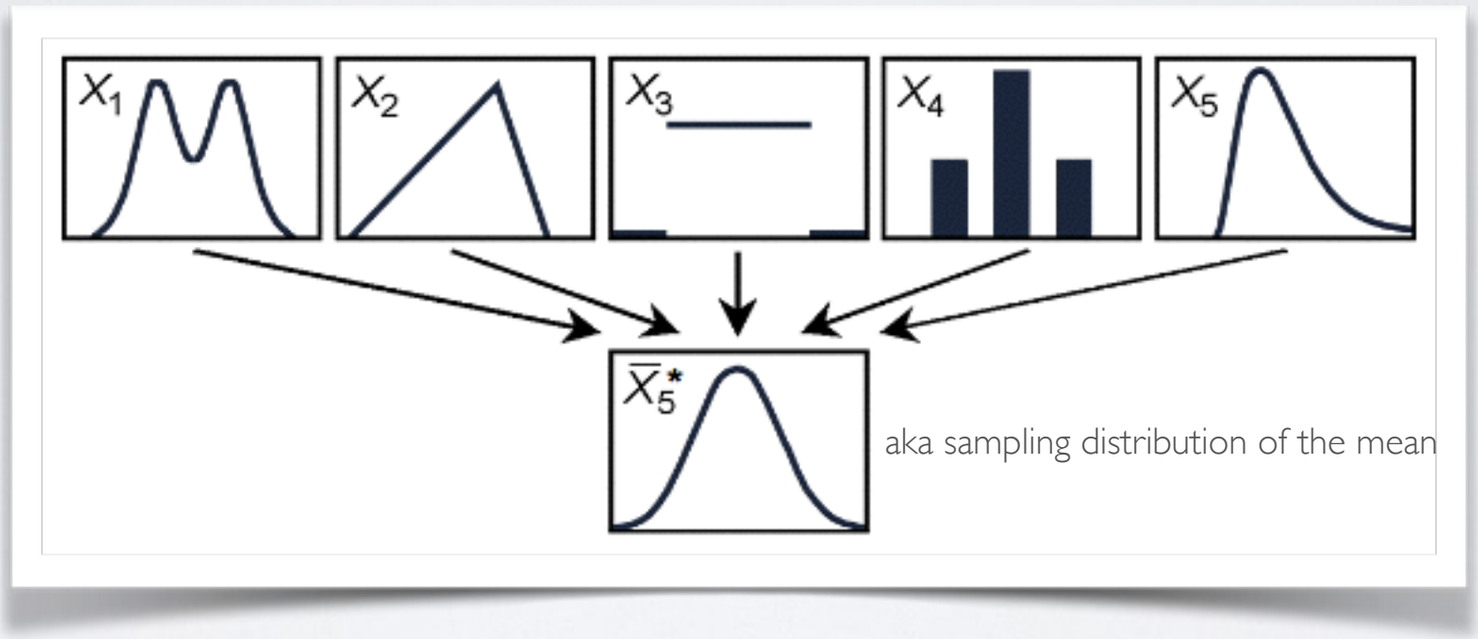
Variance



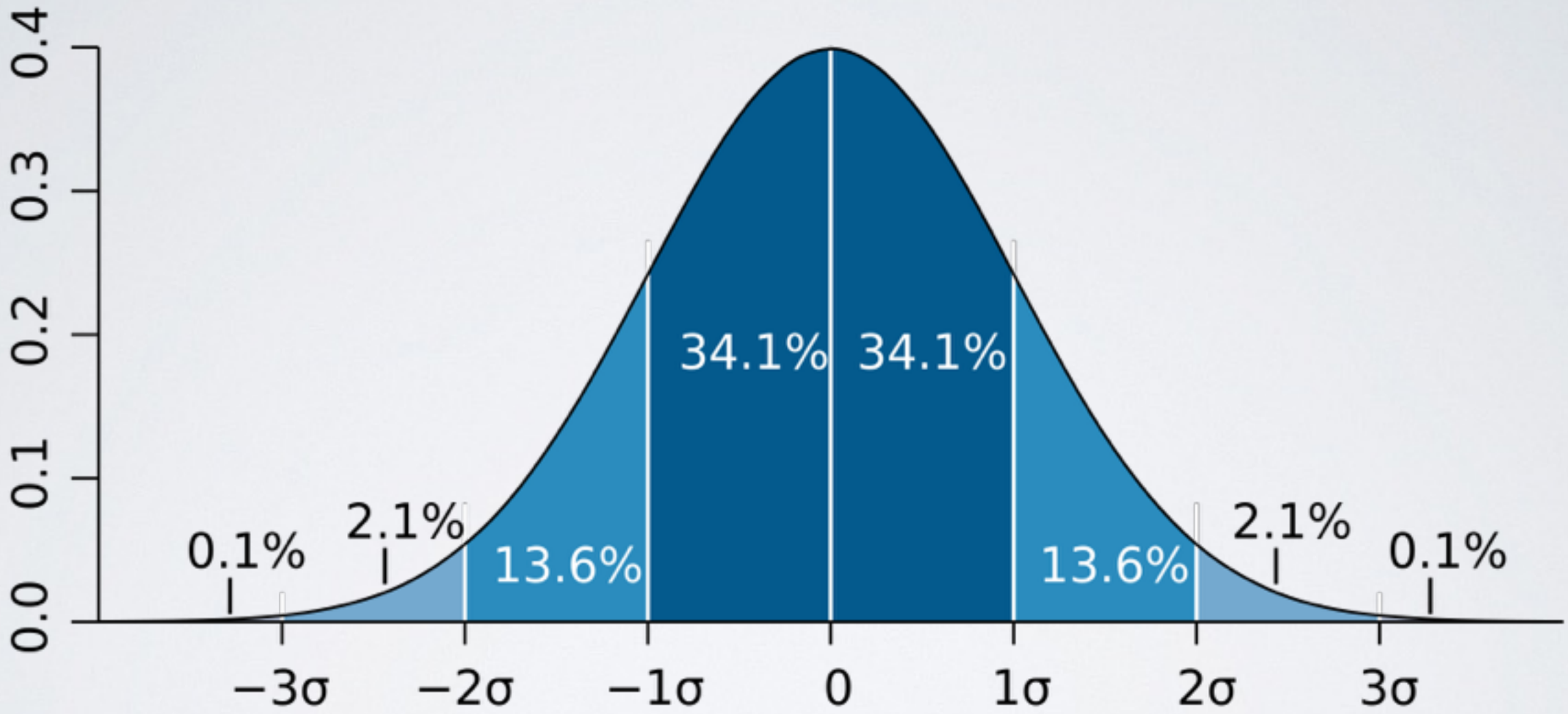
Setting a 95% confidence interval means that if you took repeated random samples from a population and calculated the statistics and CI for each sample, then the CIs for 95% of your samples would include the true value of the statistics.



Central Limit Theorem

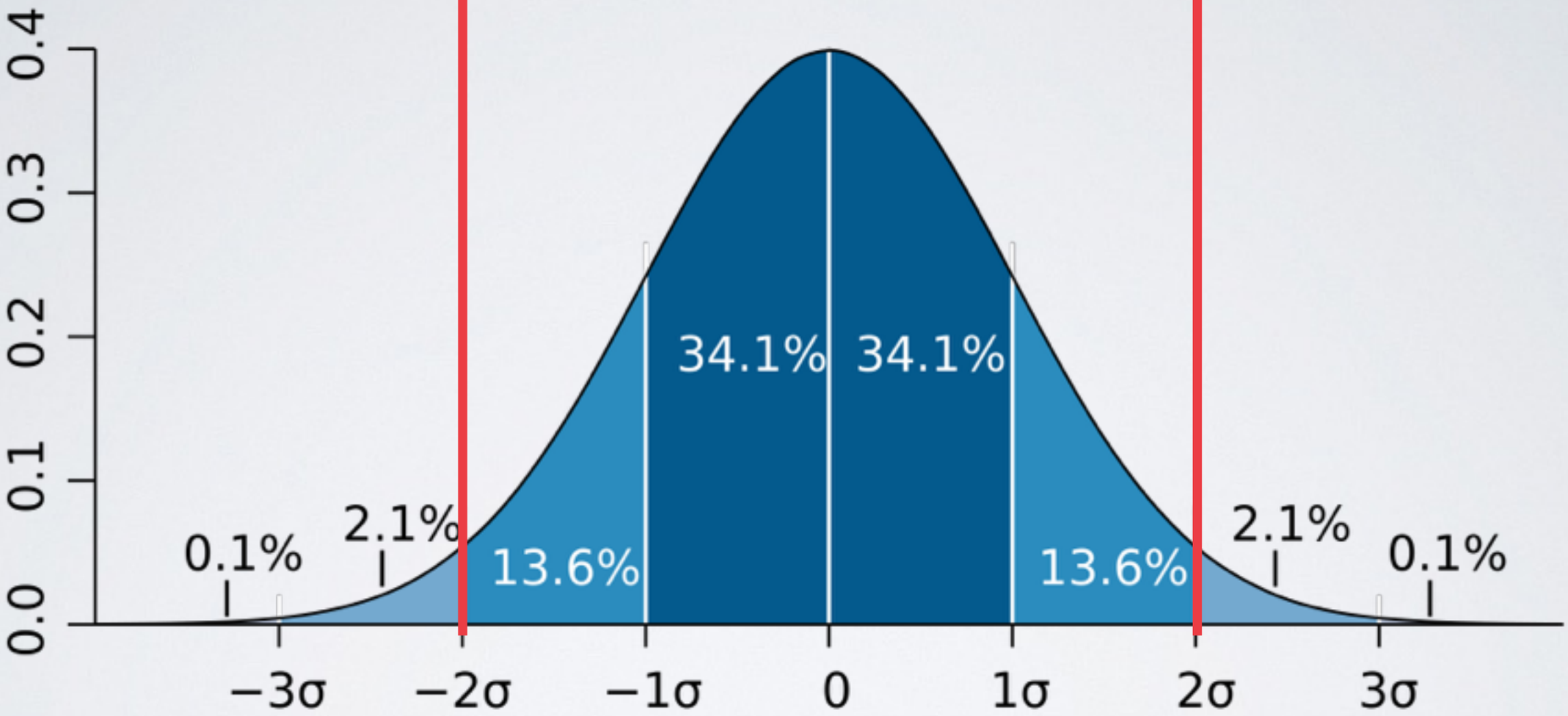


For means it's easy: the histogram of averages tends to look normal even when the histogram of the individuals doesn't!



It's easy to derive a confidence interval once we know how the theoretical sampling distribution looks like.

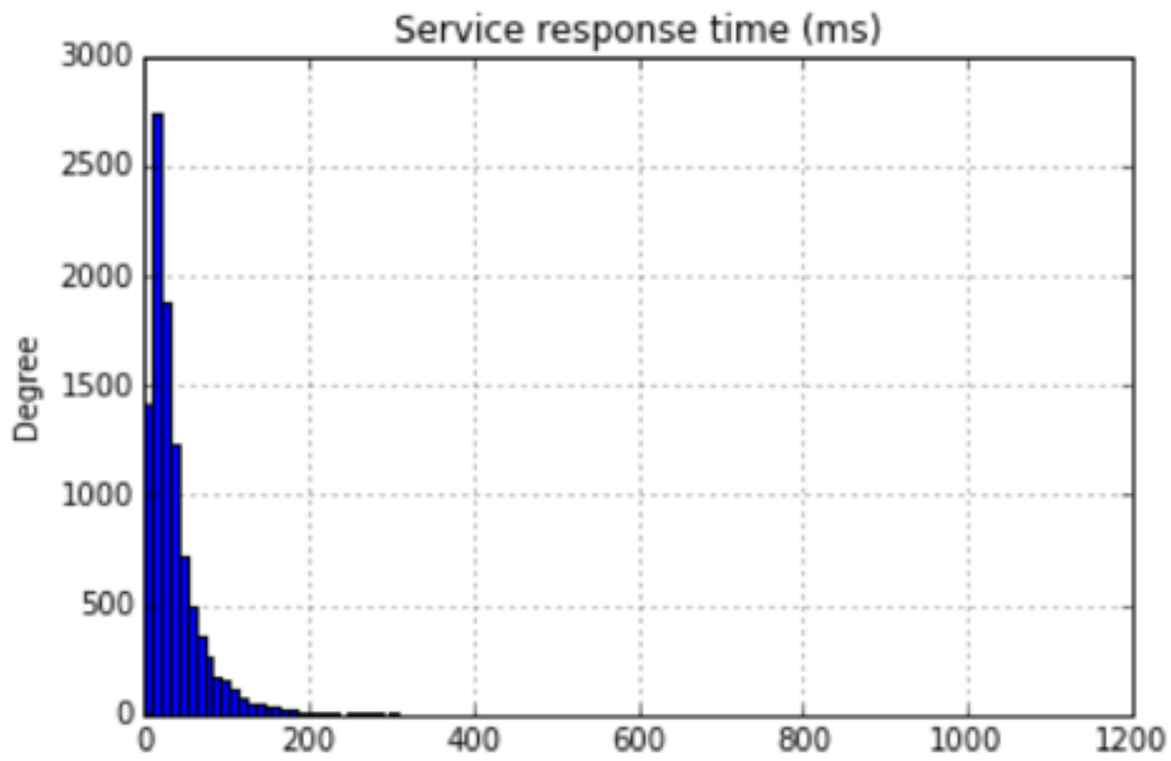
~95% confidence interval



But I don't care about means…

```
service_timings_A.plot(kind="hist", bins=100, title="Service response time (ms)")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0d424a0590>
```

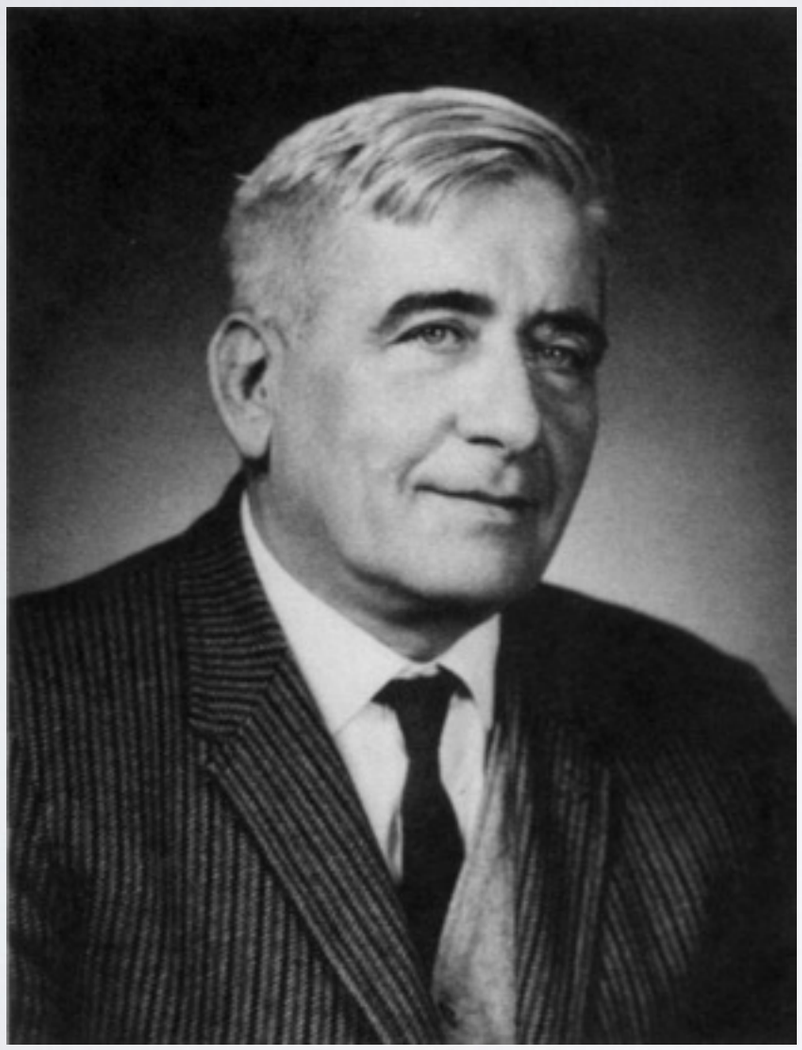



```
np.percentile(service_timings_A, 95)
```

```
109.38247299792417
```

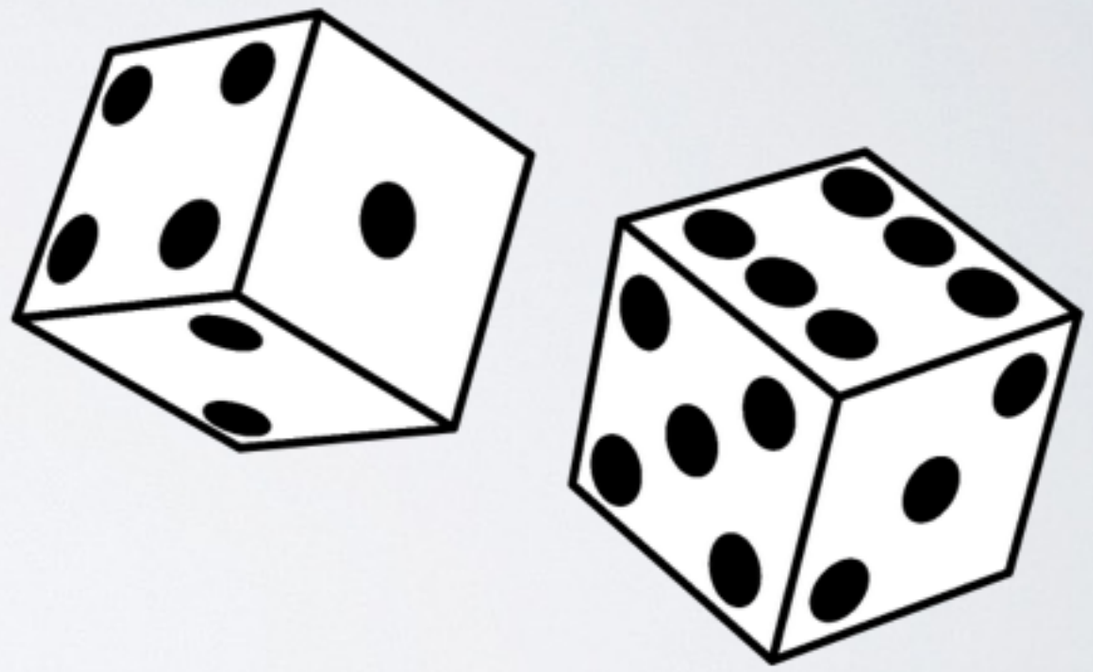
What now?

call this guy if you live in the early 20th century

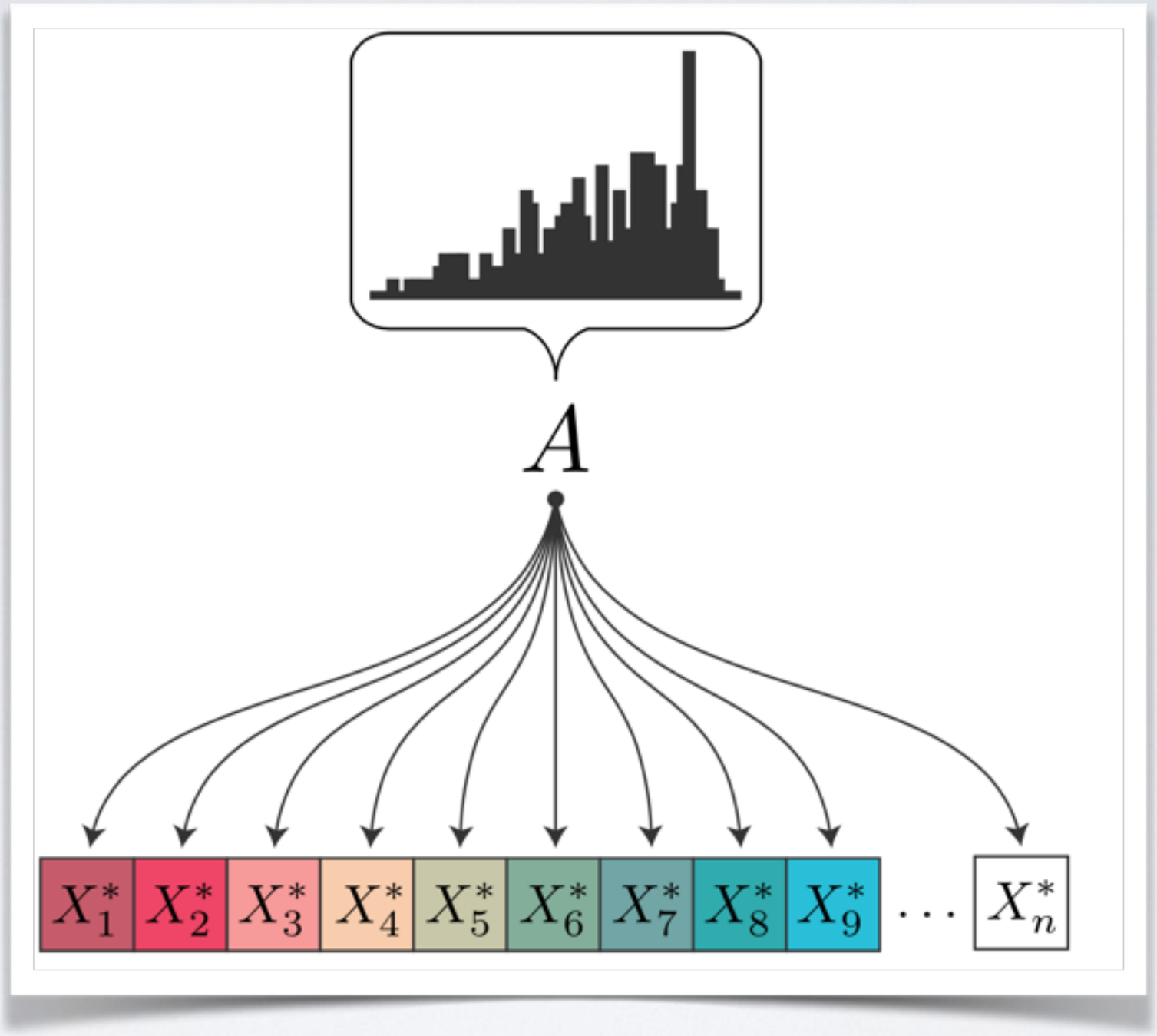


Henry Berthold Mann known for the Mann-Whitney nonparametric test

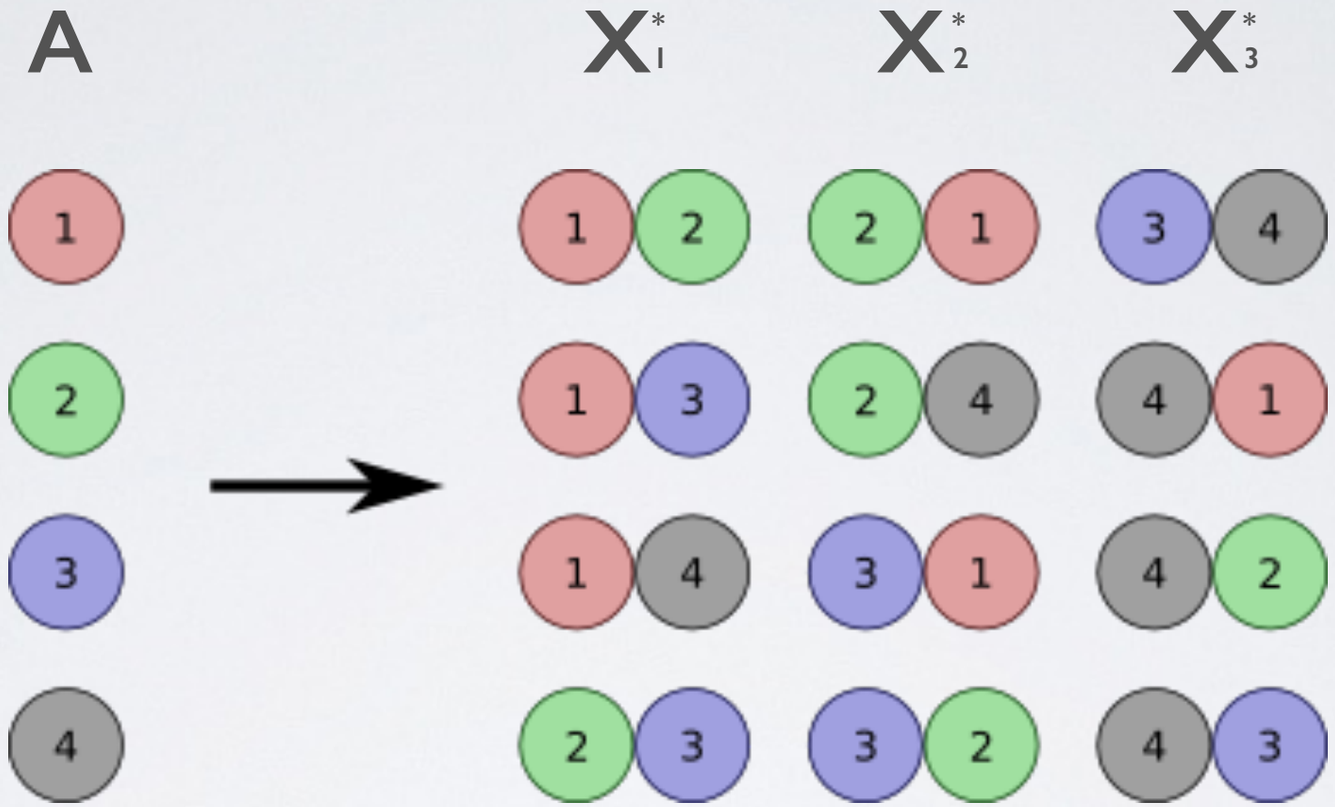
throw some (virtual) dice on your laptop



not only compilers can be bootstrapped…



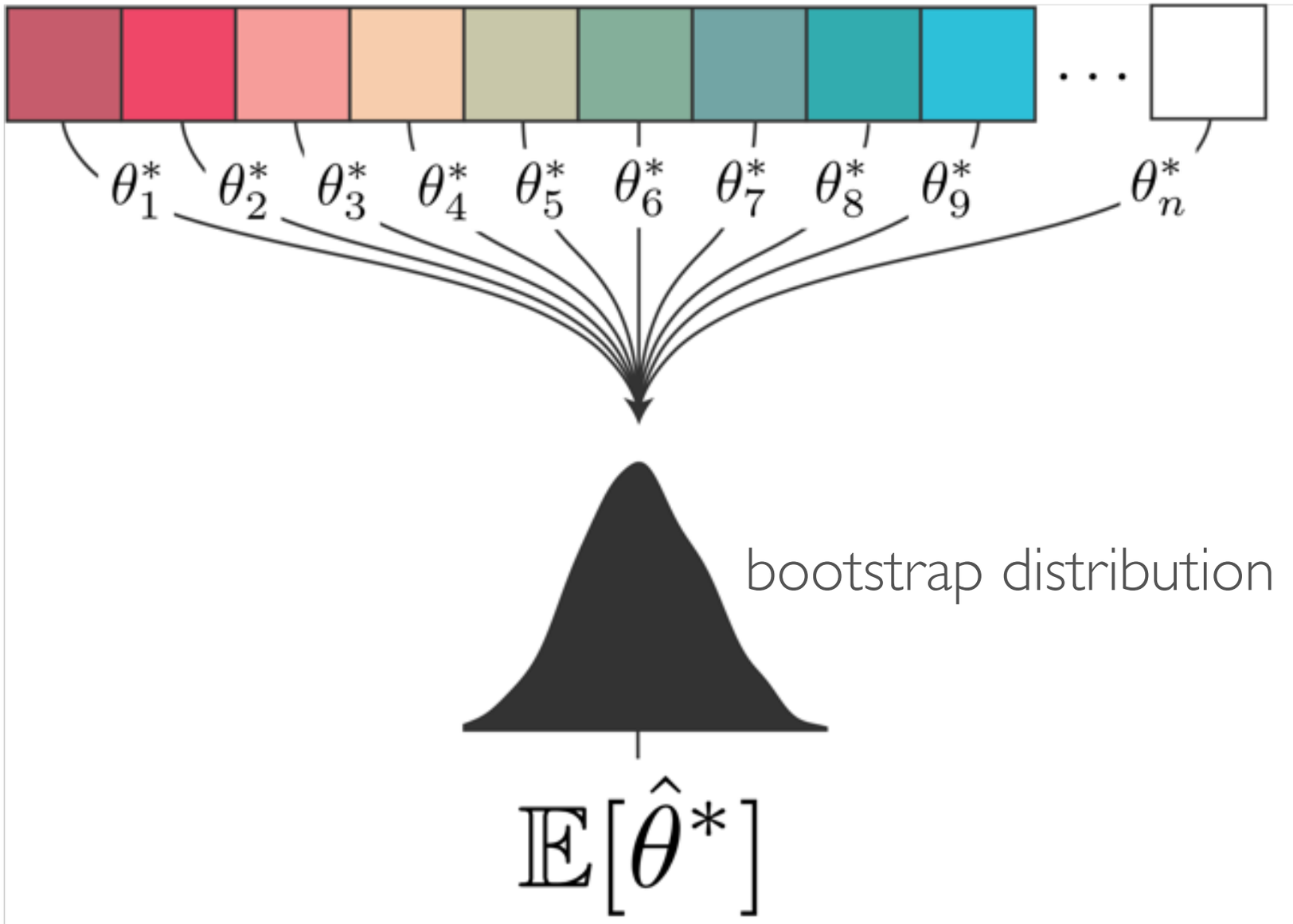
n bootstrap samples, each of size k, are generated by sampling with replacement from the original sample A


A
X^*_1
X^*_2
X^*_3
1
2
3
4
1 2
1 3
1 4
2 3
2 1
2 4
3 1
3 2
3 4
4 1
4 2
4 3

```
def bootstrap_percentile(series, percentile, n):
    bootstrap_distribution = []

    for _ in range(n):
        bootstrap_sample = resample(series)  # Sample with replacement
        #...
```

In the next phase, a bootstrap statistic is calculated for all the bootstrap samples

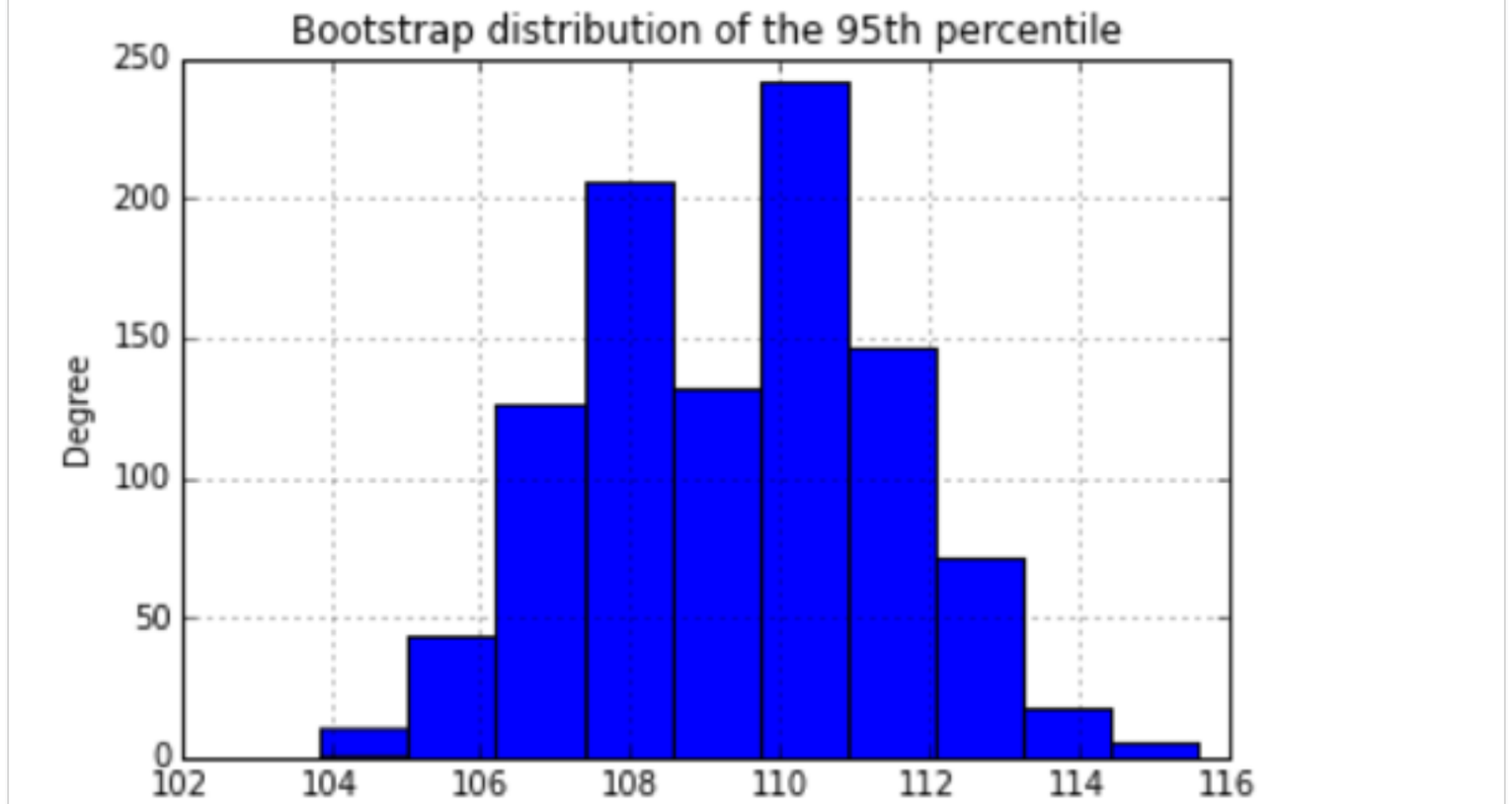


The bootstrap distribution is an approximation of the sampling distribution.

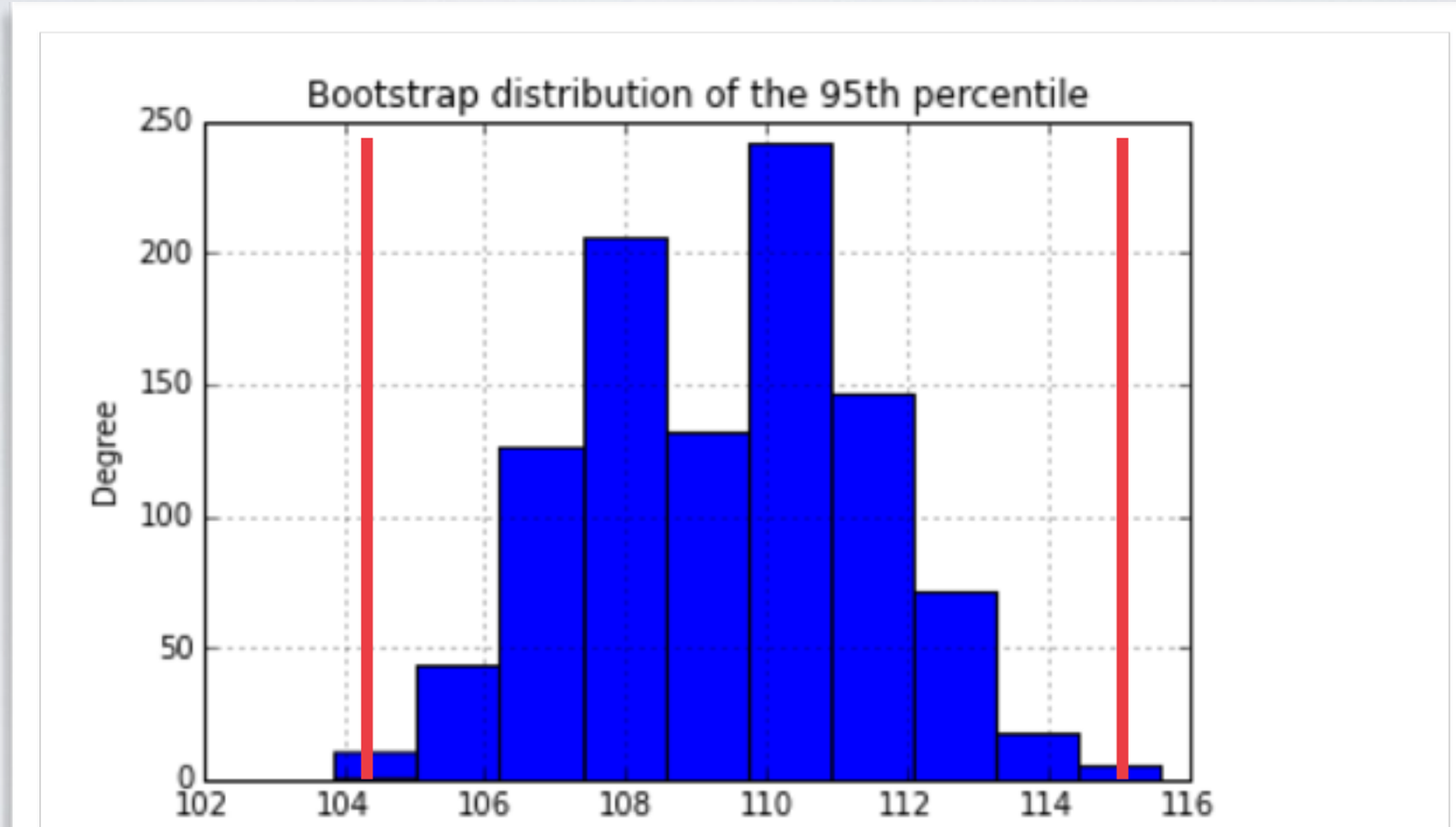
```
def bootstrap_percentile(series, percentile, n):
    bootstrap_distribution = []

    for _ in range(n):
        bootstrap_sample = resample(series)  # Sample with replacement
        statistics = np.percentile(bootstrap_sample, percentile)
        bootstrap_distribution.append(statistics)

    return pd.Series(bootstrap_distribution)
```


Bootstrap distribution of the 95th percentile
Degree
250
200
150
100
50
0
102
104
106
108
110
112
114
116

~95% confidence interval



- Resampling methods are powerful tools
- A similar procedure can be applied for A/B tests
- Checkout montecarlino