The thirst for information has made the financial industry an early adopter of data and information technology:

- Real-time data provider: Bloomberg (since 1981) and Wind (万得).
- Historical data provider: Datastream (since 1967).
- Research oriented database: CRSP (since 1960), COMPUSTAT, TAQ, etc.

Finance is about risk and uncertainty:

- Theory: modeling random events in financial markets.
- Data: historical experiences of random events.
- Empirical estimation: where models meet data.

Today, we will focus on two examples:

- Normal distribution and empirical distribution.
- Estimating the expected return $\mu = E(R_t)$. 
Where to Get Data

Wharton Research Data Services (WRDS)

<table>
<thead>
<tr>
<th>Your Subscriptions</th>
<th>Not Subscribed</th>
<th>Your Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank Regulatory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blockholders</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBOE Indexes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COMPSTAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COMPSTAT Trial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRSP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CUSIP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DMEF Academic Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dow Jones</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factset Trial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fama French &amp; Liquidity Factors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Federal Reserve Bank</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSIO online</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IBES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IHS Global Insight</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Markit Trial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mergent FISD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFLINKS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Option Metrics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Option Metrics Trial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTC Markets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Penn World Tables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHLX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEC Order Execution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAQ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thomson Reuters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRACE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WRDS SEC Analytics Suite Trial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zacks Trial</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Kenneth R. French:

Biography
Curriculum Vitae
Working Papers
Data Library
Consulting Relationships
Fama / French Forum
Contact Information
Computing Realized Stock Returns

- For a publicly traded firm, we can get
  - its stock price \( P_t \) at the end of year \( t \).
  - its cash dividend \( D_t \) paid during year \( t \).
- At the end of year \( t \), we calculate the **realized** return on the stock:

\[
R_t = \frac{P_t + D_t - P_{t-1}}{P_{t-1}} = \frac{P_t - P_{t-1}}{P_{t-1}} + \frac{D_t}{P_{t-1}}
\]

- Returns = capital gains yield + dividend yield.
- For the US markets, the best place to get reliable and clean holding-period returns is CRSP. I have applied a WRDS account for our class, which gives you access to CRSP.
For any financial instrument, the single most important number is its expected return.

Suppose right now we are in year $t$, let $R_{t+1}$ denote the stock return to be realized next year. Our investment decision relies on the expectation:

$$
\mu = E(R_{t+1}).
$$

Just to emphasize, $\mu$ is a number, while $R_{t+1}$ is a random variable, drawn from a distribution with mean $\mu$ and standard deviation $\sigma$.

To estimate this number $\mu$ with precision is the biggest headache in Finance.
Estimating the Expected Return $\mu$

- We estimate $\mu$ by using historical data:
  $\hat{\mu} = \frac{1}{N} \sum_{t=1}^{N} R_t$.

- It is as simple as taking a sample average.
- Why can this sample average of past realized returns help us form an expectation of the future?
- Because our assumption that history repeats itself. Each $R_t$ in the past was drawn from an identical distribution with mean $\mu$ and standard deviation $\sigma$. 
Time Series of Annual Stock Returns

Annual Stock Returns (in Percent) from 1927 through 2015

- 57% in 1927
- 45% in 1930
- 50% in 1950
- 45% in 1970
- 38% in 2015
- -29% in 1940
- -35% in 1950
- -44% in 1930
- -28% in 1970
- -37% in 2015
Scenarios and Their Likelihood

Learning from History: Possible Events and Their Occurrence

Number of Occurrence

Scenarios of Possible Annual Returns

1931: -44%
2008: -37%
1937: -35%
1930: -29%
1974: -28%
1935: 45%
1958: 45%
1928: 39%
1945: 39%
1975: 38%
1933: 57%
1954: 50%
Probability Distribution of a Random Event

Probability Density Function (PDF): Model vs. Data

- Mean = 12%
- Std Dev = 20%

Scenarios of Possible Annual Returns

- Normal Distribution (Model)
- Empirical Distribution (Data)
We use historical returns to estimate the number $\mu$:

$$\hat{\mu} = \frac{1}{N} \sum_{t=1}^{N} R_t$$

Recall that $R_t$ is a random variable, drawn every year from a distribution with mean $\mu$ and standard deviation $\sigma$.

As a result, $\hat{\mu}$ inherits the randomness from $R_t$. In other word, it is not really a number: $\text{var}(\hat{\mu})$ is not zero.

If this variance $\text{var}(\hat{\mu})$ is large, then the estimator is noisy.
The Standard Error of $\hat{\mu}$

- Let's first calculate $\text{var}(\hat{\mu})$:

$$\text{var}\left(\frac{1}{N} \sum_{t=1}^{N} R_t\right) = \frac{1}{N^2} \sum_{t=1}^{N} \text{var}(R_t) = \frac{1}{N^2} \times N \times \sigma^2 = \frac{1}{N} \sigma^2$$

- The **standard error** of $\hat{\mu}$ is the same as $\text{std}(\hat{\mu})$:

$$\text{standard error} = \frac{\text{std}(R_t)}{\sqrt{N}} = \frac{\sigma}{\sqrt{N}}$$
Estimating $\mu$ for the US Aggregate Stock Market

- Using annual data from 1927 to 2014, we have 88 data points.
- The sample average is $\text{avg}(R) = 12\%$. The sample standard deviation is $\text{std}(R) = 20\%$.
- The **standard error** of $\hat{\mu}$:
  \[
  \text{s.e.} = \frac{\text{std}(R)}{\sqrt{N}} = \frac{20\%}{\sqrt{88}} = 2.13\%
  \]
- The 95% confidence interval of our estimator:
  \[
  [12\% - 1.96 \times 2.13\%, 12\% + 1.96 \times 2.13\%] = [7.8\%, 16.2\%]
  \]
- The **t-stat** of this estimator is (signal-to-noise ratio),
  \[
  \text{t-stat} = \frac{\text{avg}(R)}{\text{std}(R)/\sqrt{N}} = \frac{12\%}{2.13\%} = 5.63.
  \]
The Distributions of $R_t$ and $\hat{\mu}$

\[ \hat{\mu} = \frac{1}{N} \sum_{t=1}^{N} R_t \]
Time Series of Monthly Stock Returns

Monthly Aggregate Stock Returns (in percent) from Jan 1927 through Dec 2011

mean = 0.91%

std = 5.46%
Estimating $\mu$ Using Monthly Returns

- Since the standard error of $\hat{\mu}$ depends on the number of observations, why don’t we use monthly returns to improve on our precision?
- Using monthly aggregate stock returns from January 1927 through December 2011, we have 1020 months. So $N=1020$!
- The mean of the time series is 0.91%, and std is 5.46%.
- So the standard error of $\hat{\mu}$ is:
  \[
  \text{s.e.} = \frac{5.46\%}{\sqrt{1020}} = 0.1718\%
  \]
- The signal-to-noise ratio:
  \[
  \text{t-stat} = \frac{0.91\%}{0.1718\%} = 5.30
  \]
- We increased $N$ by a factor of 12. Yet, the t-stat remains more or less the same as before.
### US and China Stock Returns

<table>
<thead>
<tr>
<th></th>
<th>CRSP VW</th>
<th>CN All</th>
<th>CN LG</th>
<th>CN Med</th>
<th>CN SM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Monthly Returns 1993-2018</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu$</td>
<td>0.83</td>
<td>1.16</td>
<td>0.99</td>
<td>1.41</td>
<td>2.02</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>4.23</td>
<td>11.05</td>
<td>10.56</td>
<td>12.49</td>
<td>13.74</td>
</tr>
</tbody>
</table>

|                |         |        |       |        |       |
| **Monthly Returns 2000-2018** |         |        |       |        |       |
| $\mu$          | 0.52    | 0.86   | 0.80  | 1.02   | 1.43  |
| $\sigma$       | 4.33    | 8.16   | 7.96  | 9.60   | 10.42 |

|                |         |        |       |        |       |
| **Monthly Returns 2010-2018** |         |        |       |        |       |
| $\mu$          | 1.00    | 0.28   | 0.21  | 0.40   | 0.99  |
| $\sigma$       | 3.71    | 6.59   | 6.38  | 8.60   | 9.59  |
The US and China Correlation in Equity Markets

EWMA Correlation using 5-Day Returns

<table>
<thead>
<tr>
<th>Year</th>
<th>Correlation (%)</th>
<th>t-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>43</td>
<td>5.25</td>
</tr>
<tr>
<td>2011</td>
<td>37</td>
<td>3.35</td>
</tr>
<tr>
<td>2012</td>
<td>28</td>
<td>3.20</td>
</tr>
<tr>
<td>2013</td>
<td>39</td>
<td>2.75</td>
</tr>
<tr>
<td>2014</td>
<td>-11</td>
<td>-0.84</td>
</tr>
<tr>
<td>2015</td>
<td>35</td>
<td>1.51</td>
</tr>
<tr>
<td>2016</td>
<td>26</td>
<td>1.73</td>
</tr>
<tr>
<td>2017</td>
<td>-7</td>
<td>-0.51</td>
</tr>
<tr>
<td>2018</td>
<td>44</td>
<td>4.83</td>
</tr>
<tr>
<td>2019</td>
<td>40</td>
<td>3.28</td>
</tr>
<tr>
<td>2020</td>
<td>44</td>
<td>4.54</td>
</tr>
</tbody>
</table>

*Rolling windows of 5-day returns; t-stats corrected for serial correlations.*
The Main Takeaways

- The financial industry has always been data intensive:
  - Data contains information.
  - Data contains noise.

- A good practitioner knows how to extract signal from noise:
  - Knowing how to read tables with standard errors and t-stats is essential.
  - Basic econometrics and statistics will be an important differentiator.

- Questions to be answered by Wednesday’s student presentations:
  - What are the means and standard deviations of monthly returns on the US and Chinese equity markets?
  - What is the correlation between the monthly returns?
  - How accurate are these estimates?