时间序列分析与预测

第一讲

黄嘉平

深圳大学｜中国经济特区研究中心
粤海校区汇文楼办公楼 1510
课程网站 https://huangjp.com/TSAF/

1. 导论

1.1. 什么是预测

《封神演义》第十一回

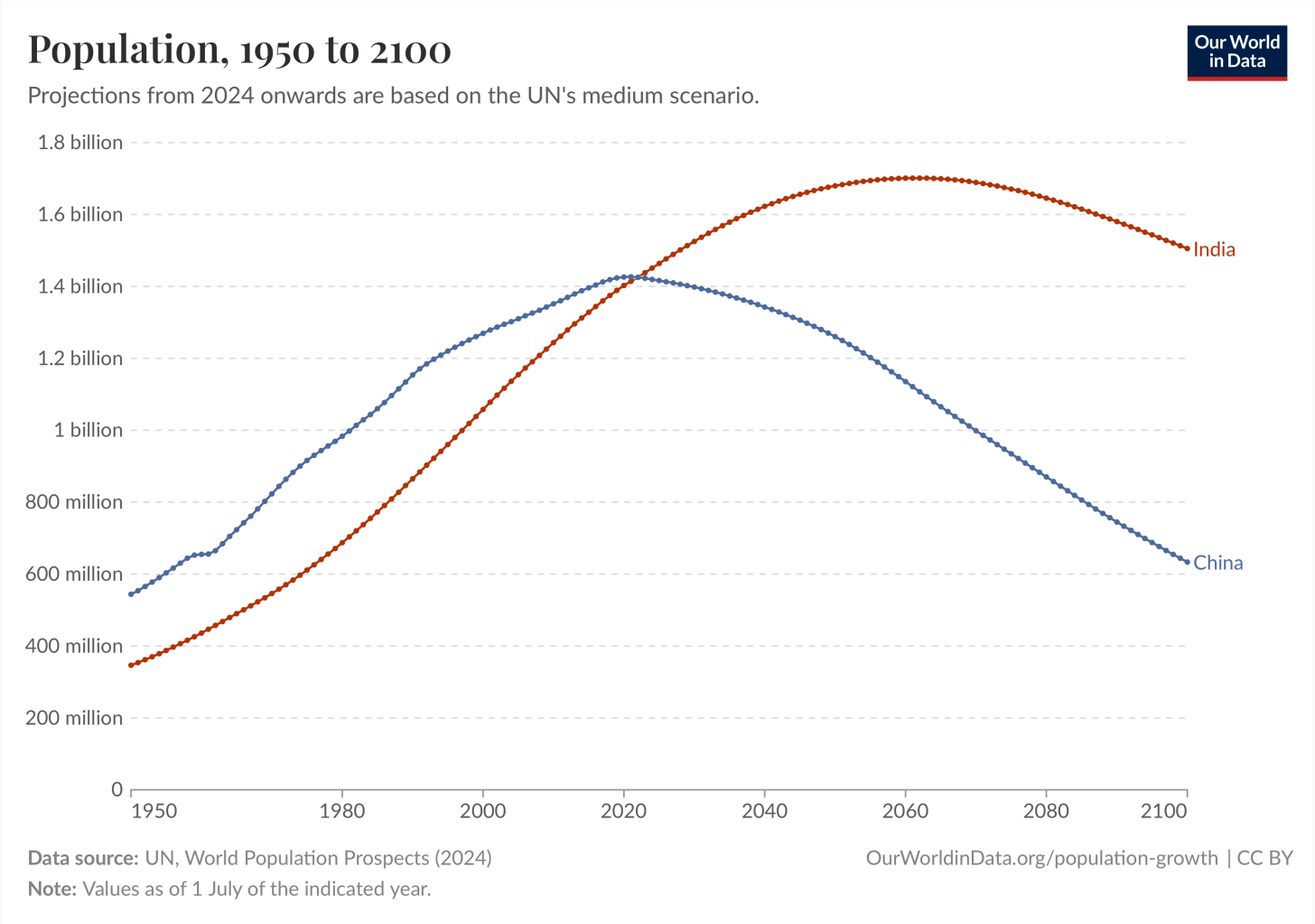
费仲曰：“请问贤侯，仲常闻贤侯能演先天数，其应果否无差？”姬昌答曰：“阴阳之理，自有定数，岂得无准？但人能反此以作善趋避之，亦能逃越。”……

少顷，二人（费仲、尤浑）又问曰：“不才二人，亦求贤侯一数，看我等终身何如？”姬伯原是贤人君子，那知虚伪，即袖演一数，便沉吟良久，曰：“此数甚奇甚怪！”费、尤二人笑问曰：“如何？不才二人数内有甚奇怪？”昌曰：“人之死生，虽有定数，或瘫痨鼓膈，百般杂症，或五刑水火，绳缢跌扑，非命而已。不似二位大夫，死得蹊蹊跷跷，古古怪怪。”

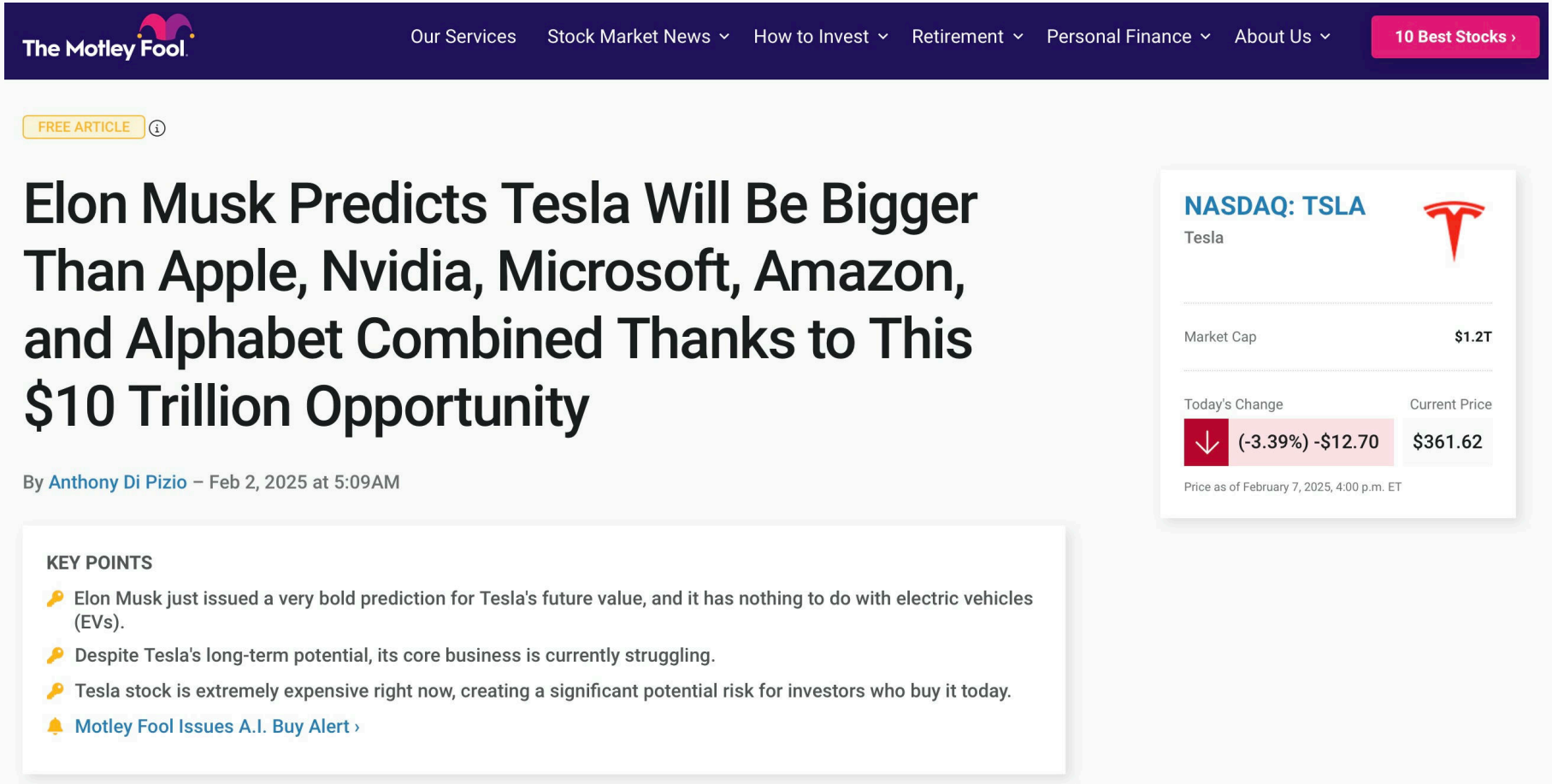
1.1. 什么是预测

msn.cn, 2025 年 2 月 10 日

1.1. 什么是预测


Population, 1950 to 2100
Our World in Data
Projections from 2024 onwards are based on the UN's medium scenario.
1.8 billion
1.6 billion
1.4 billion
1.2 billion
1 billion
800 million
600 million
400 million
200 million
0
1950
1980
2000
2020
2040
2060
2080
2100
India
China
Data source: UN, World Population Prospects (2024)
Note: Values as of 1 July of the indicated year.
OurWorldinData.org/population-growth | CC BY

1.1. 什么是预测



The Motley Fool

Our Services | Stock Market News | How to Invest | Retirement | Personal Finance | About Us | 10 Best Stocks ›

FREE ARTICLE

Elon Musk Predicts Tesla Will Be Bigger Than Apple, Nvidia, Microsoft, Amazon, and Alphabet Combined Thanks to This $10 Trillion Opportunity

By Anthony Di Pizio – Feb 2, 2025 at 5:09AM

KEY POINTS

- Elon Musk just issued a very bold prediction for Tesla's future value, and it has nothing to do with electric vehicles (EVs).
- Despite Tesla's long-term potential, its core business is currently struggling.
- Tesla stock is extremely expensive right now, creating a significant potential risk for investors who buy it today.
- Motley Fool Issues A.I. Buy Alert ›

NASDAQ: TSLA
Tesla

Market Cap	$1.2T
Today's Change	(-3.39%) -$12.70
Current Price	$361.62

Price as of February 7, 2025, 4:00 p.m. ET

https://www.fool.com/investing/2025/02/02/elon-musk-tesla-bigger-apple-nvidia-10-trillion/

1.1. 什么是预测

预测 (forecast/forecasting) 在不同资料中的定义

- **Cambridge Advanced Learner's Dictionary & Thesaurus (online)**
  forecast [verb]: to say what you expect to happen in the future.
- **Merriam-Webster's Dictionary (online)**
  forecast [verb]: to calculate or predict (some future event or condition) usually as a result of study and analysis of available pertinent data.
- **Oxford Dictionary of Statistics**
  forecasting: the prediction of future values in a time series.
- **中国大百科全书（第三版网络版）**【所属学科：安全科学与工程】
  预测：预测即由已知推知未来，包含了两个方面的含义：(1) 根据过去已有的历史资料和当前的实际情况，使用科学的理论和方法，分析和推测未来可能出现的突发事件情况；(2) 估计和推测已知突发事件的未来发展变化。

1.1. 什么是预测

书中 1.2 节对预测 (forecasting) 的描述：

预测指在利用所有可用信息（包括可能对预测结果产生影响的历史数据和知识）的情况下尽可能准确地预言未来。

Forecasting is about predicting the future as accurately as possible, given all of the information available, including historical data and knowledge of any future events that might impact the forecasts.

1.2. 预测对象和效果

每个领域都有自己关心的预测对象，例如每年国家都会提出经济发展目标，其依据是针对宏观经济运行情况的预测（主要指标是 GDP）。

在管理学领域，我们可能更关心和企业发展相关的因素。例如：

- 电力公司需要预测中短期的天气变化用以估计电力需求
- 国际贸易企业需要预测汇率和国际政治局势的变化趋势
- 零售行业需要预测每种商品的需求变化，以决定补充库存的时间和数量（便利店模式）

在实践中，预测的效果取决于你对预测对象的了解程度，有多少数据可以利用，是否有明确的趋势性特征，以及预测结果能否影响对象的发展趋势。

1.3. 如何进行预测

1. **确定预测目标**（例：电力公司预测天气变化）
   - 变量是什么？（例：气温、降雨概率）
   - 需要预测的时间跨度和频次？（例：两周内、日度）
   - 有哪些可用数据？（例：过去一年的日度气象数据）
2. **选择预测方法**

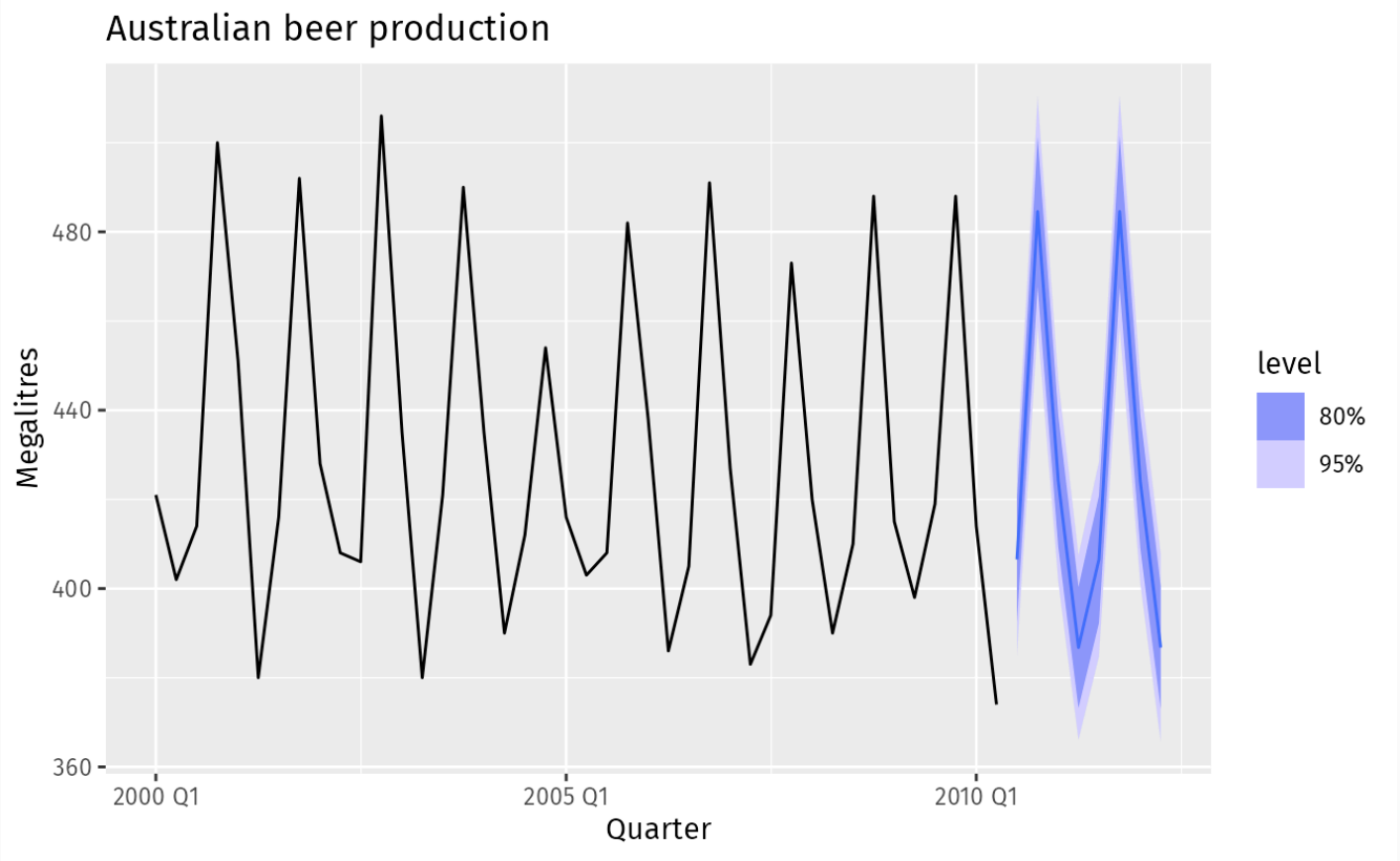
   - 定性方法：适用于数据缺失时
     例：假设你打算新开一个线下体育用品商店，那么你要如何决定首批商品的进货量？

   - 定量方法：适用于历史数据可用，且一些特征保持不变时
     - 横截面数据：多个个体在同一时间点的数据（例：2024 年各省 GDP）
     - **时间序列数据**：同一个体在多个时间点的数据（例：广东省 1990-2024 年度 GDP）

1.4. 时间序列数据的预测

我们只涉及定期观测的数据（每小时、每天、每周、每月、每季度、每年等）。

时间序列数据预测的目的是**估计将来某时间点上该序列的取值**，例如右图展示了 2000-Q1 至 2010-Q2 澳大利亚的季度啤酒产量，以及对未来两年的预测。



1.5. 预测的主要步骤

预测通常包含以下五个基本步骤

1. **定义问题**
2. **收集信息**
   不仅要收集统计数据，数据收集者和预测者的经验积累也很重要
3. **初步（探索性）分析**
   根据数据绘图，并尝试从中发现规律
4. **选择并拟合模型**
   常用模型包括回归模型、指数平滑法、Box-Jenkins ARIMA 模型、动态回归模型等
5. **使用和评估模型**

1.6. 预测的统计学视角

我们要预测的事情是未知的，因此是**不确定**的。在统计学中，我们将不确定的量称为随机变量（random variable）。

一般情况下，需要预测的时间点离现在越近，不确定性程度就越低，预测精确度就越高；离现在越远则不确定性程度越高，预测精确度越低。

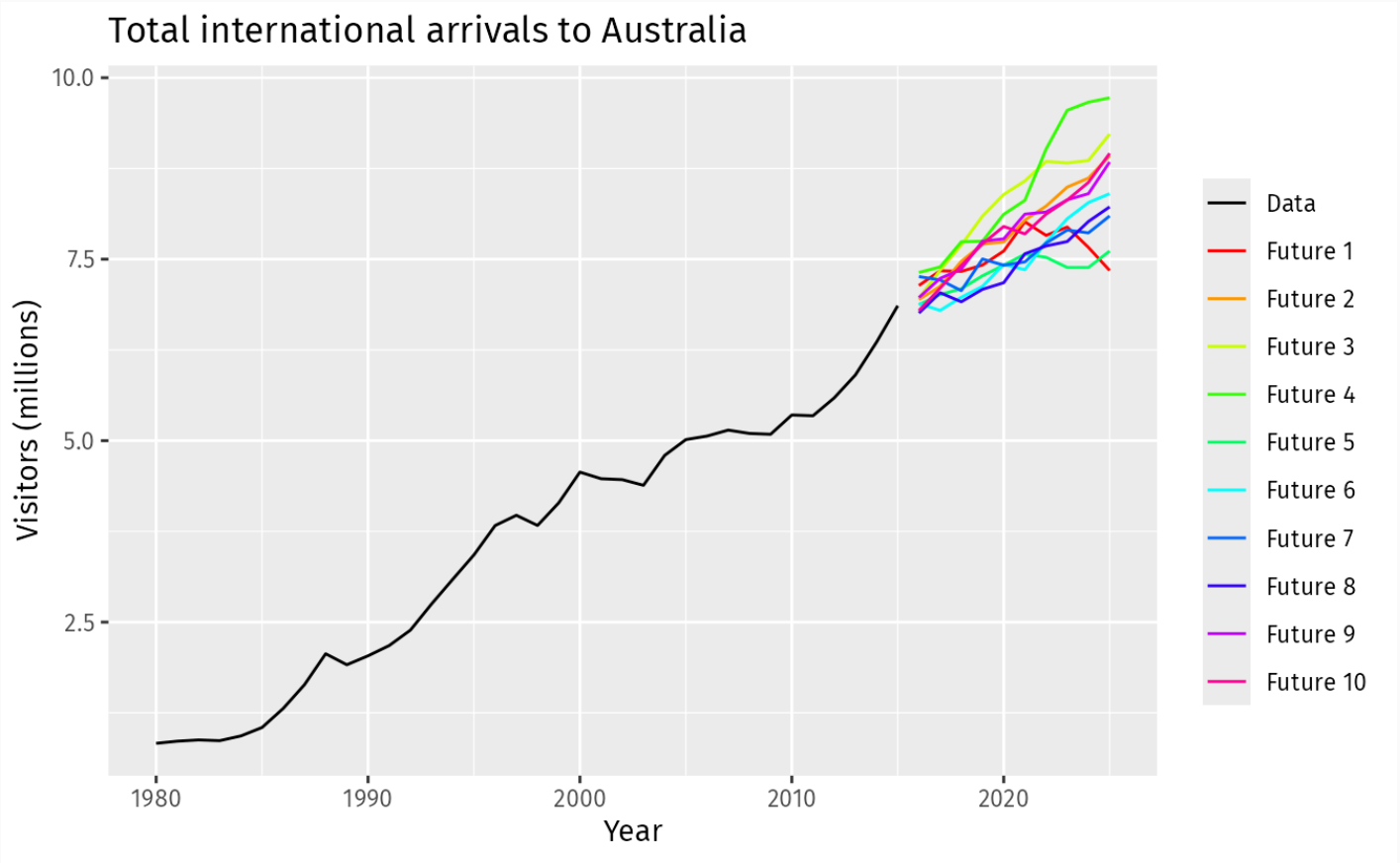
这里介绍一些相关的数学表达

- 在时间 t 的随机变量写作 y_t。
- 已知信息的集合写作 $\mathscr{I}$ (手写体的 I，在不同字体下形状也不同)。
- 基于 $\mathscr{I}$ 的随机变量 y_t 写作 $y_t|\mathscr{I}$，其分布称为预测分布。
- 统计学上的“预测值”通常指的是预测分布的期望值，写作 $\hat{y}_t$。
- $\hat{y}_{t|t-1}$ 代表在 $t-1$ 时预测的 y_t，也就是利用了观测值 $(y_1, \ldots, y_{t-1})$，类似的，$\hat{y}_{T+h|T}$ 代表在 T 时预测的 h 期后的 y_{T+h}。

1.6. 预测的统计学视角

在基于数据和模型的定量预测中，利用拟合后的模型可以人为的“随机”生成不同的未来发展路径，每一条路径都是一种预测结果。

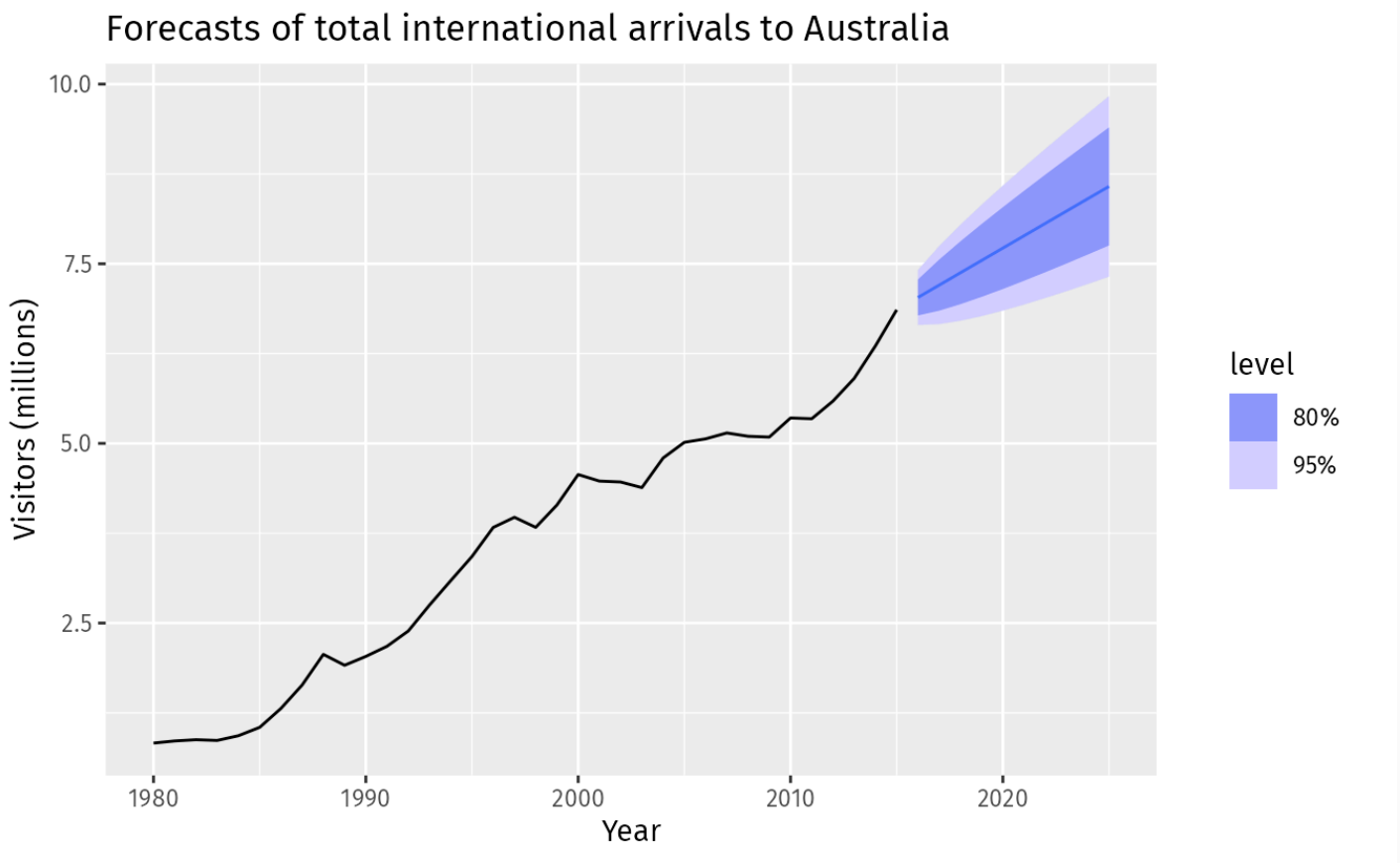
右图展示了澳大利亚年度外国访客数以及十种可能发生的未来路径。



1.6. 预测的统计学视角

获得所有未来路径是不可行的。我们更习惯于给出**预测区间**（prediction interval）。一个 95% 预测区间意味着真实值被包含在该区间内的概率是 95%。

如果你只需要一个预测值，那么可以用所有可能的未来路径的均值，这通常称为**点预测**。



2. R 的基础用法

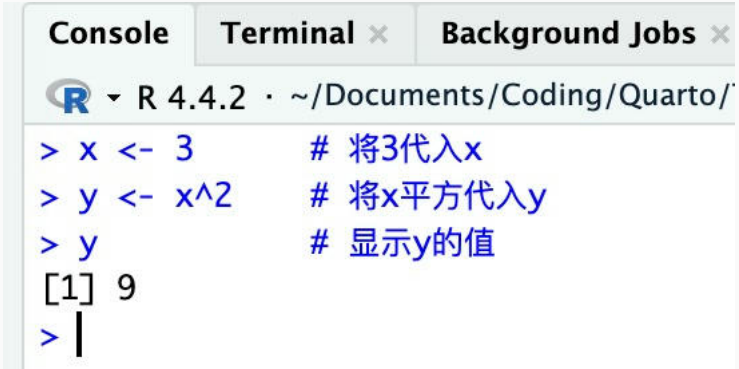
2.1. R 和 RStudio 简介

- R 是著名的免费开源统计计算和绘图软件，同时也是一种编程语言。
  https://www.r-project.org/
- RStudio 是为 R 设计的综合开发系统（IDE, integrated development enviriment）。
  https://posit.co/products/open-source/rstudio/
- R 可以在所有操作系统下运行，你可以用任意的文本编辑器进行 R 的编程，并在任意的命令行软件下执行。但配合 RStudio 使用可以大幅提升工作效率和体验。

  如果没有 R，RStudio 本身什么也做不了。如果把 RStudio 比做电脑，R 就相当于一个操作系统，没有安装操作系统的电脑对于普通用户来说没有任何用处 s。
- 安装 R 和 RStudio：
  访问 https://posit.co/download/rstudio-desktop/ 并根据提示依次安装 R 和 RStudio

2.2. R 程序与运行结果的显示方式

R 程序在 RStudio 中的显示方式及运行结果：



本课件中程序和运行结果的显示方式：

```
x <- 3        # 将 3 代入 x
y <- x^2      # 将 x 平方代入 y
y             # 显示 y 的值
#> [1] 9
```

程序部份（前三行）每行中 # 后面的部份是注释，执行时会被忽略。

2.3. Packages

刚刚安装好的 R 仅包含最基础的功能（可称为 base R），就像刚刚安装好的 Windows 操作系统，虽然蕴含无限可能，但功能有限或不太好用。

我们购买新电脑后的第一件事或许就是安装 Office/WPS 等常用的软件。与此相似，R 也需要附加工具包（package）的支持才能发挥更大的作用。工具包通常包含 base R 中没有的命令和一些数据集，需要先**安装**，然后在每次重新启动 R（RStudio）时进行**调用**才能使用。这和每次重启电脑后都要打开 Office 才能进行文档编辑是类似的。

R 官方支持的工具包数量超过 2 万个，还有很多发表在 Github 上，但每个领域常用的工具包并不多。在本门课程的学习中，我们基本上只需要调用一个工具包 `fpp3`。这是教材作者整合了一些热门工具包和数据集后重新打包制作的。

工具包的安装和调用方法请参考 **课程网站 > 课程资料 > 第一部份：导论和基础知识**

2.4. 零基础学习 R

以下内容参考 Torfs, P., & Brauer, C. (2024). A (very) short introduction to R.

R 的最基本功能是计算（当作计算器使用）。例如计算 $10^2 + 36$：

```
10^2 + 36
#> [1] 136
```

可将数值代入变量后再进行调用或计算：

```
a <- 4      # 代入（"<-" 和 "=" 都可以代入数值，更推荐使用 "<-"）
a           # 显示变量内容
#> [1] 4
```

可以利用快捷键 Alt + - （Mac 上是 ⌥ + -）快速输入代入符号 <- 。

```r
a * 5    # 计算 a 乘以 5
#> [1] 20
```

```r
a = a + 10     # 更新 a 的内容
a              # 显示 a 的内容
#> [1] 14
```

练习：计算你的 BMI（搜索 BMI 的定义，并代入自身数值进行计算）。

2.4. 零基础学习 R

R 中保存数值的基本方式分为标量（scalar）、向量（vector）和矩阵（matrix）。

向量就是数值的一维列表（称为 array），可以用 `c()` 命令生成，例如：

```
b = c(3,4,5)  # 将 (3,4,5) 作为向量代入 b
b             # 显示 b 的内容
#> [1] 3 4 5
```

调取向量中的单个数值：

```
b[2]          # 显示 b 中的第二个数值
#> [1] 4
```

2.4. 零基础学习 R

在使用 R 的过程中，最常用的不是简单计算，而是 base R 或工具包中包含的各种函数（function）。

如果我们要计算 3, 4, 5 三个数值的均值，可以直接计算

```
(3 + 4 + 5) / 3
#> [1] 4
```

也可以用 `mean()` 函数

```
mean(b)
#> [1] 4
```

`help()` 是查询其他函数帮助文档的函数，运行 `help(rnorm)` 了解 `rnorm()` 的作用。同样的功能也可以用 `?rnorm` 实现。

矩阵在 R 中等同于 2 维列表。我们可以用 `matrix()` 函数生成向量

```
mat <- matrix(data = c(9,2,3,4,5,6), ncol = 3)
```

注意函数 `matrix()` 包含 `data` 和 `ncol` 两个参数。`data` 指定矩阵的内容，`ncol` 指定矩阵的列数（类似的还有 `nrow`）。

这里我们将 $9, 2, 3, 4, 5, 6$ 这六个数字以 2×3 矩阵的形式存入了变量 `mat`。默认的排列顺序是先横向后纵向。

```
mat
```

```
#>      [,1] [,2] [,3]
#> [1,]    9    3    5
#> [2,]    2    4    6
```

提取第 2 行用 `mat[2,]`，提取 1 列用 `mat[,1]`，提取第 $(2, 1)$ 要素用 `mat[2,1]` 。

R 中最基本的数据保存形式是 data frame。它和矩阵很像，但是每一列都有对应的名称，这也是保存数据和保存数值的本质区别。

我们在保存数据时，通常让每一列代表一个变量，而每一行代表一个观测值。Data frame 基本体现了这种习惯。例如小李、小王和小张三人的身高和体重可以保存为

```
body_data <- data.frame(Name = c("小李", "小王", "小张"), Height =
c(175, 172, 181), Weight = c(70, 68, 82))
```

```
body_data
```

```
  Name Height Weight
1 小李    175     70
2 小王    172     68
3 小张    181     82
```

调取 data frame 中的内容时，除了指定要素的具体位置以外，也可以通过指定列名称的方式调取一整列数值。下面两种方法都能调取 `body_data` 中的身高

```
body_data[,2]
```

```
[1] 175 172 181
```

```
body_data$Height
```

```
[1] 175 172 181
```

如果想要调取小张所属的那一行，可以采用条件判断的方式

```
body_data[body_data$Name == "小张",]
```

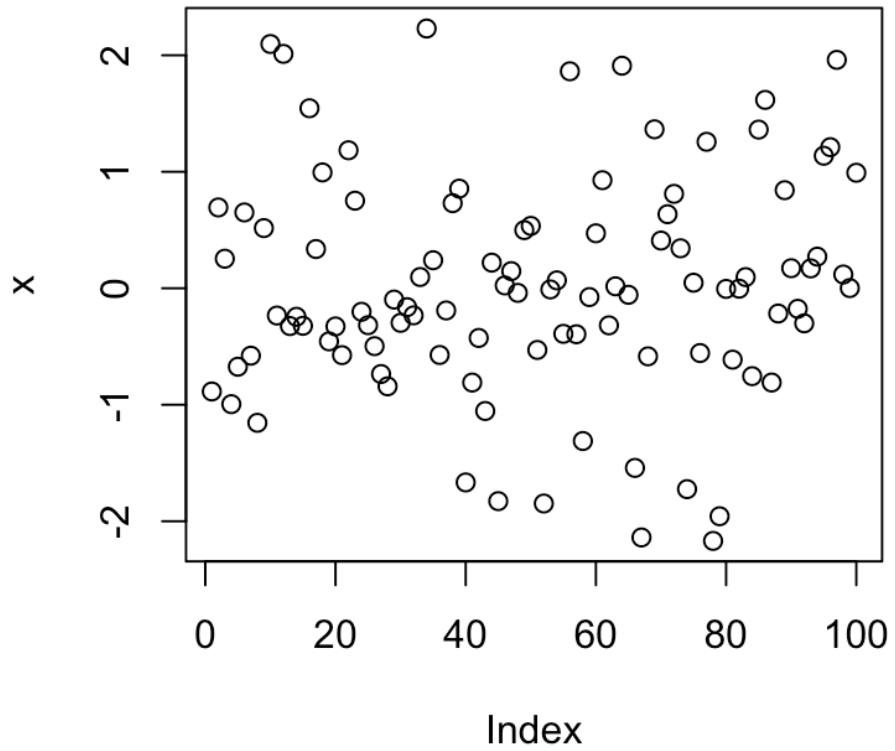
```
  Name Height Weight
3 小张    181     82
```

2.4. 零基础学习 R

R 的另一个强大功能是绘图。这里首先介绍 base R 的绘图命令，后面会改用 `ggplot2` 工具包绘制更加精美的图表。

```
x = rnorm(100)
  # 你知道上面一行在做什么吗?
plot(x)
```

绘制的图表会显示在 RStudio 右下方的 Plots 面板中



2.4. 零基础学习 R

Base R 中的常用函数

- 数据生成：`c()`, `array()`, `matrix()`, `data.frame()`, `rbind()`, `cbind()`
- 调取数值：`x[n]`, `x[m,n]`, `x[c(k,m,n)]`, `x[x>m & x<n]`, `x$n`
- 获取变量的信息：`length()`, `ncol()`, `class()`, `print()`
- 统计相关：`sum()`, `mean()`, `sd()`, `max()`, `min()`, `rowSum()`, `quantile()`
- 数据处理：`seq()`, `rnorm()`, `sort()`, `t()`, `cumsum()`, `paste()`
- 模型拟合：`lm(v1 ~ v2)`, `coef()`, `summary()`
- 基本绘图：`plot(x)`, `plot(x,y)`, `hist()`, `barplot()`, `abline()`

可以通过 `help()` 了解每个函数的用法。但更重要的是一定要自己动手尝试运行，只去看是无法学会任何一个编程语言的。

3. 课后练习

3. 课后练习

1. 运行下面的命令并了解命令 `a:b` 的含义（`a` 和 `b` 可以是任意数值）

```
1:20
```

2. 用 `help()` 了解 `hist()` 函数的作用，并运行下面的命令

```
hist(rnorm(200))
```

   结果符合你的预期吗?

3. 用 R 计算 $4.5, 6.6, 5.6, 4.0, 4.3, 5.4, 2.6, 5.9, 5.1, 2.9$ 这十个数值的样本均值和样本方差。尝试使用公式和调用函数两种方式并比较结果。