week 1

怎样用统计学获得数据支持

- Established a research question, or hypothesis, that can be tested.
- Determined some relevant variables.
- Identified our population of interest.
- Gathered some data by taking a sample from the population.
- Analysed our data and the relevant variables.
- Formed an inference or conclusion regarding the original hypothesis

Concepts

population mean μ population variation σ^2

sample mean $\bar{X}$ sample variation s^2

数据的分类

- Categorical Data
  - Nomial, 各种无关联形容词
  - Ordinal, poor/fair/good
- Numeric

    - 离散
    - 连续

Descriptive tools for Categorical Data

mode, frequency, bar chart, pie chart

ordinal 只是多一个 order

Descriptive tools for Numeric Data

- Mean, median, mode.
- Quantiles.
- Range, interquartile range, variance, coefficient of variation.
- Covariance, correlation.
- Histograms.
- Boxplots.
- central trendency

relative standing

measure 某个数据在总体的位置，比如 quatiles, 31st percentile

percentile

$L_p = (n+1)\frac{p}{100}$ 数据有8个(n=8)，那么 31st percentil(L_{31}) 在

$L_{31} = (8+1)\frac{31}{100} = 2.79$

interquartile range (IQR)

$IQR = Q_1 - Q_3$

Variance

population variance σ^2

$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$

sample variance s^2

$s^2 = \frac{1}{n-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$

shortcut

$$s^2 = \frac{1}{n-1}\left(\left(\sum_{i=1}^{n} X_i^2\right) - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right)$$

Standard deviation

population sd σ

sample sd s

* Coefficient of Variation

用来比较两组(很可能是规格不同的)数据的分散情况，cv越大，表示越分散

population $CV = \frac{\sigma}{\mu}$

sample $cv = \frac{s}{\bar{X}}$

covariance

协方差是两个变量之间线性关系的度量，描述它们如何相互关联

population covariance

$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^{N} (X_i - \mu_X)(Y_i - \mu_Y)$

sample covariance

$sXY = \frac{1}{n-1} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})$

shortcut

$$s_{XY} = \frac{1}{n-1}\left(\left(\sum_{i=1}^{n} X_i Y_i\right) - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}\right)$$

* Correlation Coefficient

数据相关性的标化指标，比如有 A,B 两组数据，想知道 A与C 更相关，还是 B与C更相关，就可以用这个指标

值的范围是-1.0至1.0

population cor

$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$

sample cor

$r_{XY} = \frac{s_{XY}}{s_X s_Y}$

P47的例题数据 用R计算

R

```
> # 把数据存入变量 X
> X <- c(8.3, -6.2, 20.9, -2.7, 33.6, 42.9,
24.4, 5.2, 3.1, 30.5)

> # 把另一组数据存入变量 Y
> Y <- c(12.1, -2.8, 6.4, 12.2, 27.8, 25.3,
18.2, 10.7, -1.3, 11.4)

> # average/mean
> mean_X = mean(X) # Xbar = 16

> # Standard Deviation
> sd_X = sd(X) # s = 16.74336

> # Variance
> var_X = var(X) # s^2 = 280.34

> # Coefficient of Variation
> cv_X = sd(X)/mean(X)

> # Covariation
> cov_XY = cov(X, Y)
```

```
> # Correlation Coefficient
> cor_XY = cor(X, Y)
```