Quantitative Reasoning

# Probability and inference

## KEY CONCEPTS

**Kelvin Horia**
**pvokh@nus.edu.sg**
Provost office, QR team

# Sample space and events

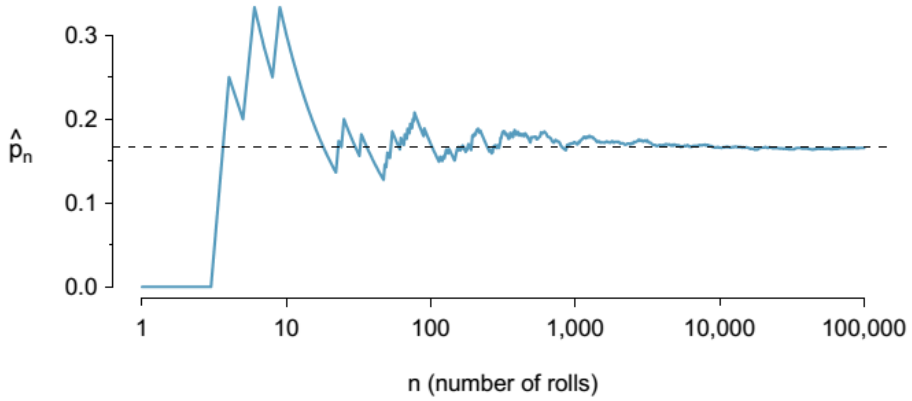
- A **sample space**, denoted by *S*, is the set of all possible outcomes of a **random process** (also known as a probability experiment)
  - **Discrete random variable** : One where the sample space of outcomes forms a discrete numerical variable.
    - Example : Throwing a dice. *S* = {1, 2, 3, 4, 5, 6}
  - **Continuous random variable** : One where the sample space of outcomes forms a continuous numerical variable
    - Example : Normally distributed IQ scores with mean 100 and variance 25.
- An **event** of the sample space is a **subset** of the sample space
  - *S* = {1, 2, 3, 4, 5, 6} with throwing the dice being a random process
    - Event *A* can be *A* = {1, 3, 5} $\subseteq S$
    - Event *B* can be *B* = {2, 6} $\subseteq S$.
  - *S* = all possible IQ scores.
    - Event *A* can be all scores between 100 to 120.

If each outcome is equally likely to occur.

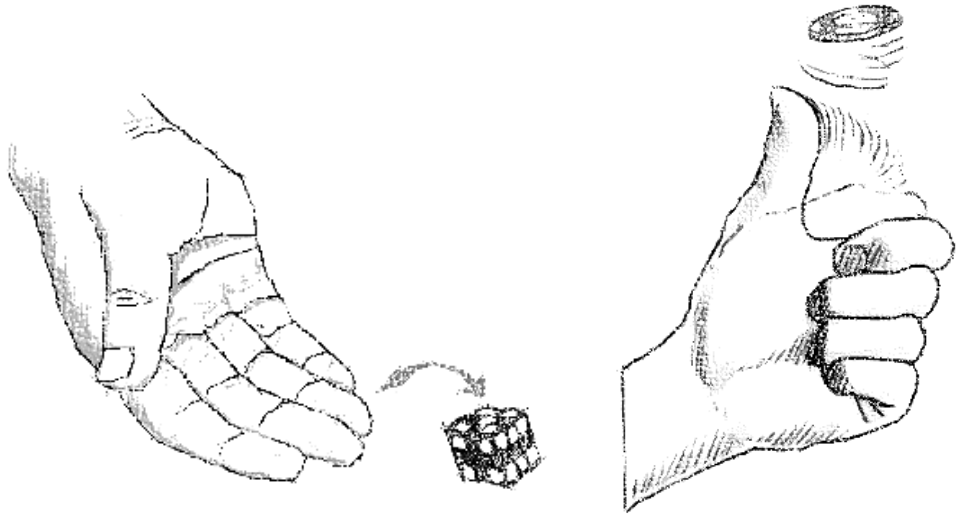
$$P(A) = \frac{\text{size of } A}{\text{size of sample space.}}$$

# Probability

- Theoretical probability vs empirical observations
  - Chance processes that can be repeated over and over again independently and under the same conditions will have empirical observations that are close to the theoretical probability.
    - Rolling a die
    - Tossing a coin




The fraction of die rolls that are 1 at each stage in a simulation. The proportion tends to get closer to the probability $1/6 \approx 0.167$ as the number of rolls increases.



# Some definitions and formulas

## Definitions

- We say events A and B are **mutually exclusive** when A and B cannot occur **at the same time.**

  Mathematical formulation : $P(A \cap B) = 0$

- We say events A and B are **independent** when the occurrence of one event DOES NOT influence the likelihood of the other event occurring.

  Mathematical formulation : $P(A|B) = P(A)$

  Chance of A happens if B already happen.

  $P(A \cap B) = P(A) \cdot P(B)$

- They are talking about 2 totally different aspects of probability and shouldn't be mixed up.

## Formulas

- Complement rule

  $P(A) = 1 - P(A^C)$

  not A

  Is useful

- Addition rule
  - For mutually exclusive events A and B

    $P(A \cup B) = P(A) + P(B)$
  - For all events A and B

    $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

    0 if mutually exclusive
- Multiplication rule
  - For Independent events A and B

    $P(A \cap B) = P(A) \cdot P(B)$
  - For all events A and B

    $P(A \cap B) = P(A) \cdot P(B|A)$

$P(A|B) = \frac{P(A \cap B)}{P(B)}$

# The link between discrete probability and rates

**Rates**

| | Heart disease | No heart disease | Row total |
|---|---|---|---|
| Smoker | 38 | 14 962 | 15 000 |
| Non-Smoker | 44 | 84 956 | 85 000 |
| Column total | 82 | 99 918 | 100 000 |

**Drawing randomly** from the pool of 100 000

Every person has the **same** chance of being chosen

**Probabilities**

| | Heart disease | No heart disease | Row total |
|---|---|---|---|
| Smoker | 38 | 14 962 | 15 000 |
| Non-Smoker | 44 | 84 956 | 85 000 |
| Column total | 82 | 99 918 | 100 000 |

Rate(Smoker) = 15%

$$\text{Rate(Heart Disease} \mid \text{Smoker)} = \frac{38}{15\,000}$$

$$\text{Rate(Heart Disease and smoker)} = \frac{38}{100\,000}$$

P(Smoker) = 15%

$$\text{P(Heart Disease} \mid \text{Smoker)} = \frac{38}{15\,000}$$

$$\text{P(Heart Disease and smoker)} = \frac{38}{100\,000}$$

- Rates can be converted to probability with the same values under random selection with every unit having same chance of being chosen

# Link between independence and association

**Rates**

| | Heart disease | No heart disease | Row total |
|---|---|---|---|
| Smoker | 38 | 14 962 | 15 000 |
| Non-Smoker | 44 | 84 956 | 85 000 |
| Column total | 82 | 99 918 | 100 000 |

**Drawing randomly** from the pool of 100 000

Every person has the **same** chance of being chosen

**Probabilities**

| | Heart disease | No heart disease | Row total |
|---|---|---|---|
| Smoker | 38 | 14 962 | 15 000 |
| Non-Smoker | 44 | 84 956 | 85 000 |
| Column total | 82 | 99 918 | 100 000 |

$$\text{Rate(Heart Disease} \mid \text{Smoker)} = \frac{38}{15\,000}$$

$$\text{Rate(Heart Disease)} = \frac{82}{100\,000}$$

$$\text{P(Heart Disease} \mid \text{Smoker)} = \frac{38}{15\,000}$$

$$\text{P(Heart Disease)} = \frac{82}{100\,000}$$

Heart disease and smoking **are associated!** → Smoking and heart disease are **not independent!**

- *A* and *B* not being associated to each other is tantamount to saying that *A* and *B* are independent of each other.

# Rules for rates translated to probabilities

**Rates**

- **Symmetry**
  - If $\text{rate}(A \mid B) \neq \text{rate}(A \mid \text{not } B)$ then $\text{rate}(B \mid A) \neq \text{rate}(B \mid \text{not } A)$
    - If $\text{rate}(A \mid B) > \text{rate}(A \mid \text{not } B)$ then $\text{rate}(B \mid A) > \text{rate}(B \mid \text{not } A)$
    - If $\text{rate}(A \mid B) < \text{rate}(A \mid \text{not } B)$ then $\text{rate}(B \mid A) < \text{rate}(B \mid \text{not } A)$
  - If $\text{rate}(A \mid B) = \text{rate}(A \mid \text{not } B)$ then $\text{rate}(B \mid A) = \text{rate}(B \mid \text{not } A)$
- **General rule**
  - Given $\text{rate}(A \mid B) = x$ and $\text{rate}(A \mid C) = y$ with *B* and *C* disjoint, we will have

$$\min\{x, y\} \leq \text{rate}(A \mid B \cup C) \leq \max\{x, y\}$$

**Probabilities**

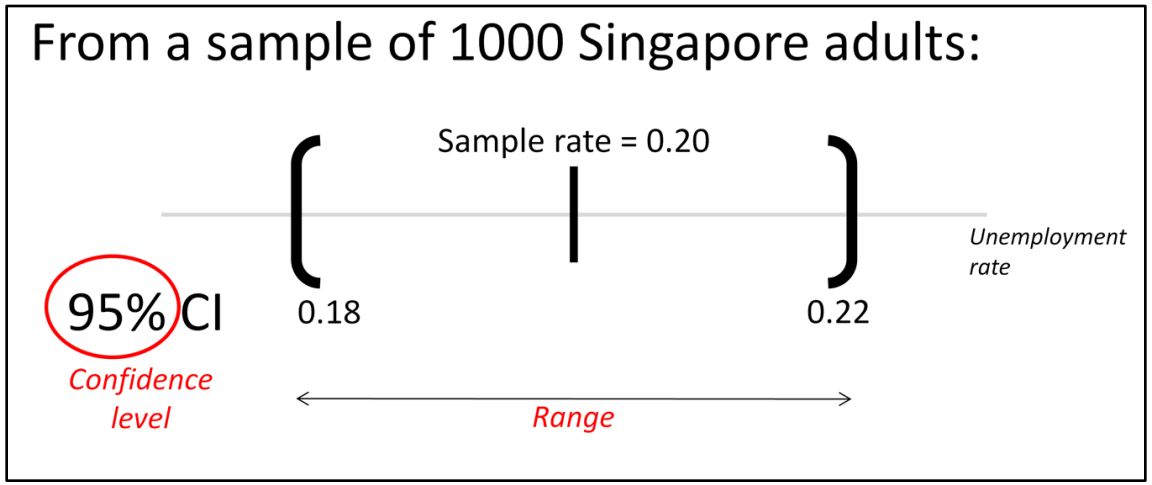
$P(A|B) > P(A|\text{not } B)$

$\Rightarrow P(B|A) > P(B|\text{not } A)$

$P(A)$ is always between $P(A|B)$ and $P(A|\text{not } B)$.

# Confidence intervals

- Suppose no bias: **Sample's estimate = Parameter + Random error**




95% CI: **0.20** ± 0.02

- We are 95% **confident** that **[0.18,0.22]** contains the population parameter.

Why can't we say chance?

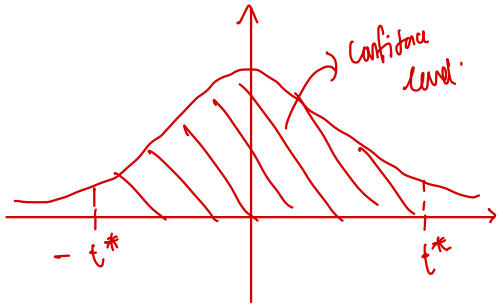
- This means if 100 samples of similar sizes are collected, via the same sampling procedure 95 of the samples' CIs would contain the population parameter

# Width and assumptions of confidence intervals

**For proportions**

$$p^* \pm z^* \times \sqrt{\frac{p^*(1-p^*)}{n}}$$

where $p^*$ denotes the sample proportion




**For means**

t- distribution with degree of freedom = n-1 / sample size.

$$\bar{x} \pm t^* \times \frac{s_x}{\sqrt{n}}$$

where $\bar{x}$ denotes the sample mean

- What's the trade-off if one wishes to have a confidence interval with a high level of confidence? (say 99%)
- From either of the formulas above, how will sample size affect the width of the confidence interval?
- Suppose you were working with the entire population data. Based on either of the formulas above, you can still input the values and calculate an "interval". What does this interval mean?

For same sample
if confidence level ↗ .
width of CI ↗ .

# Hypothesis Testing (the general strategy)

- We observe a phenomena and ask ourselves whether it's due to chance or not.

- So we come up with 2 hypothesis namely the Null Hypothesis and Alternate Hypothesis
  - Null hypothesis takes a stance of no difference or no effect.
  - Alternate hypothesis is typically what we wish to confirm and pit against the null hypothesis.

- We set a level of significance (typically 0.05 or 0.01)

- To determine whether the observation is due to chance, we calculate the probability of obtaining a test result at least as extreme as the result observed **assuming Null is True(called the *p*-value). Note: extreme is interpreted as "favorable to the alternative hypothesis".**

- If *p*-value is lower than significance level, we can reject the null hypothesis.

# A scenario for hypothesis testing

- **Claim**: A coin manufacturer claims that he/she has produced a biased coin with P(H) = 0.3 and P(T) = 0.7. Imagine you toss this coin 8 times and you obtained heads on the first 7 tosses and tails on the last toss.

**Question that comes to mind**: Can the above observation happen due to chance? Or **is the coin manufacturer's claim not correct?** To answer this question, you can do a **hypothesis test**.

Hypothesis test

**Null hypothesis**: Coin is as claimed. i.e P(H) = 0.3

**Alternative hypothesis**: P(H) > 0.3 (one-tailed test)

**P-value**

$$p = (0.3)^7(0.7) + 7(0.3)^7(0.7) + (0.3)^8$$
$$= 1.29 \times 10^{-3} < 0.05$$

Since p < 0.05, we can reject the null hypothesis. In layman terms, this means that the chances of obtaining 7H and 1 Tails as a pure fluke is **so low** that we can treat it as close to impossible. We can conclude that P(H) > 0.3 which means the coin manufacturer's claim is not correct.

- Note : If p not low enough, cannot reject null hypothesis which means **you don't know whether the observation is due to chance or not.** Not rejecting the null hypothesis **doesn't** mean the null hypothesis is true.

- Note : There is no scenario in which we attempt to reject the alternate hypothesis since *p*-value can't determine anything about the alternate hypothesis.

# Example: One sample t-test

- Researcher A claims that the average resale price of HDB is 600 000 SGD while researcher B claims that the average resale price of HDB is bigger than 600 000 SGD.
- We draw a simple random sample of 1000 HDB and we find the sample average=690 000
- **Question:** Can the observation above happen due to chance?
- **Null hypothesis:** average=600 000
- **Alternative hypothesis:** average>600 000
- How do we compute the p-value? (see Tutorial 4)

# Chi-sq test

association btw 2 categorical vars.

- **Claim**: A researcher wants to test whether smoking is associated to heart disease. Assume he does a simple random sampling of 15 000 smokers and 85 000 non-smokers and observes the following.

## Observation

| | Heart disease | No heart disease | Row total |
|---|---|---|---|
| Smoker | 38 | 14 962 | 15 000 |
| Non-Smoker | 44 | 84 956 | 85 000 |
| Column total | 82 | 99 918 | 100 000 |

$$\text{Rate(Heart disease)} = \frac{82}{100\,000}$$

## Expected numbers

Null hypothesis : Smoking is not associated with heart disease. I.e

Confirm association $r(H.D \mid S) = r(H.D \mid \text{non - smoker}) = \frac{82}{100\,000}$

Alternate hypothesis : Smoking is associated with heart disease

| | Heart disease | No heart disease | Row total |
|---|---|---|---|
| Smoker | 12.3 | 14897.7 | 15 000 |
| Non-Smoker | 69.7 | 84930.3 | 85 000 |
| Column total | 82 | 99 918 | 100 000 |