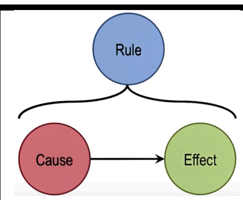
## The Scientific Method

| | |
|---|---|
| Observation | Derive from an observer's use of his senses at a particular place and time |
| | Significant condition (what to write down):<br>• Need theory to guide what factors are relevant and distinguish irrelevant ones<br>• Theory is built into even our most basic "observation statements"<br>• Unless we can eliminate some 'irrelevant' factors, we will never be able to generalize<br>• As science progresses, new theoretical terms and observational language are invented<br>• Never absolutely certain, due to some assumptions/errors/lack of theory |
| Hypothesis<br><br>Good Theory | Write down a general hypothesis that captures all these observations |
| | • Every detail of the theory is needed for the predictions (Hard to vary when wrong)<br>• Able to easily generate multiple predictions that could be easily falsified or tested<br>• Good theory may not convince everyone (VS mythical stories)<br>• Theories govern all events of a particular kind at all places and all times |
| Occam's Razor: | Unknown parameters: Measure them via experiments, use these obtained values to make more quantitative predictions, compare predictions with other experiments to confirm/falsify |
| | Have as few parameters (assumptions) as possible, and make as many predictions as possible |
| | • Complicated curve more likely to be wrong than simple one, may only fit for specific data<br>• Only use when cannot tell which hypothesis is true<br>• Describe why cannot use a simple relationship (Why can't we use a straight line?) |
| Prediction | Theory cannot be ambiguous, predictions should be as concrete as possible (use math) |
| Experiment | Test the hypothesis, see if the predictions match with the results of the experiment: |
| Confirmation<br>Falsification | Confirm: Continue to derive other consequences with the theory/make a more general theory |
| | Falsify (mismatch between theory and experiment): Return to the observational phase<br>Hypothesis need not completely discard right away (many possible explanations, lack theory) |
| | • Compare predictions from theory with experiments as much as possible<br>• Theories can only be falsified, never sure if it is 100% correct |
| Reasoning | Induction: Generalize from a large number of observation statements, under a wide variety of conditions (may be wrong) CE→R<br>Abduction: Just a proposed explanation for observation for a <u>single</u> observation (may be wrong) RE→C<br>Deduction: Find specific consequences given a law (No ambiguity) RC→E |
| Guesstimation | 1. Start with basic information (use round numbers **(round answer)**, explain the reasoning)<br>2. Make reasonable assumptions 3. Break down problem into manageable parts |
| | Estimate answer within a factor of 10 \| Geometric mean: $\sqrt{\max \times \min}$ |
| Modelling | An idealized representation or example for thinking about a real-life phenomenon<br>Express our hypothesis in terms of equations and numerical parameters, make predictions |
| | • Identify important aspects, Make appropriate assumptions<br>• Decide on key quantities that model should describe, Forget about non-essential factors |




## Difference Equations

Each term is defined as a function of preceding term(s)

| O | H | Type | Solution (closed form) | |
|---|---|---|---|---|
| 1 | N | $x_n = x_{n-1} + b$ | $x_n = x_0 + nb$ (A.S.) | |
| 1 | Y | $x_n = rx_{n-1}$ | $x_n = r^n x_0$ | H: No const. or $n$ |
| 1 | N | $x_n = rx_{n-1} + b$ | $x_n = r^n\left(x_0 + \frac{b}{r-1}\right) - \frac{b}{r-1}$ | |
| 2 | Y | $x_n = Ax_{n-1} + Bx_{n-2}$ | | |

| | | |
|---|---|---|
| General Sol.<br>Particular Sol. | $x_n - x_{n-1} - 2x_{n-2} = 0$<br>$\lambda^2 - \lambda - 2 = 0$<br>$\lambda = -1 \; or \; \lambda = 2$<br>$x_n = A(-1)^n + B(2)^n$ | $x_n = 4x_{n-1} - 4x_{n-2}$<br>$\lambda^2 - 4\lambda + 4 = 0$<br>$\lambda = 2 \; (Double \; root)$<br>$x_n = A(2)^n + Bn(2)^n$ |

## Differential Equations

| Separable: | 2nd Order Linear Homogeneous: |
|---|---|
| $\frac{dy}{dx} = \frac{g(x)}{f(y)}, f(y) \neq 0$<br>$f(y)\,dy = g(x)\,dx$<br>$\int f(y)\,dy = \int g(x)\,dx$<br>$F(y) = G(x) + C$<br>$y = \cdots$<br><br>Sub question to solve $C$ | $y'' - y' - 2y = 0$<br>$\lambda^2 - \lambda - 2 = 0$<br>$\lambda = -1 \; or \; \lambda = 2$<br>$y = Ae^{(-1)x} + Be^{(2)x}$<br><br>$y'' = 4y' - 4y$<br>$\lambda^2 - 4\lambda + 4 = 0$<br>$\lambda = 2 \; (Double \; root)$<br>$y = Ae^{2x} + Bxe^{2x}$ |
| Singular solution: $y \equiv 0$<br>When $\deg(f) < 0$ | General solution, put in initial conditions to solve for $A, B$ |

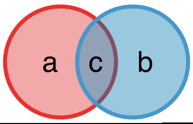
# Probability and Statistics

## Probability

$A = \{a, c\}, B = \{b, c\}, \bar{A} = \{b\}$

$A \cup B = \{a, b, c\}, \bar{A} \cup \bar{B} = \{a, b\}, A \cap B = \{c\}$

**Additive law**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$




Exam: Draw tree diagram/table!
If all else fails, use nCr/nPr
Remember order of actions!

$P(\bar{A} \cap \bar{B}) = 1 - P(A \cup B) = 1 - [P(A) + P(B) - P(A \cap B)]$

P (both girls | at least 1 girl)

$= \frac{P(G_1 \text{ and } G_2)}{P((G_1 \text{ and } G_2) \text{ OR } (B_1 \text{ and } G_2) \text{ OR } (G_1 \text{ and } B_2))}$

P (both G | elder child G)

$= \frac{P(G_1 \text{ and } G_2)}{P((G_1 \text{ and } G_2) \text{ OR } (B_1 \text{ and } G_2))}$

P(no match)=1-P(at least one match)= $1 - \frac{365}{365} \times \frac{364}{365} \times \ldots \times \frac{365-n+1}{365}$

P(no match with you) = $1 - \left(\frac{364}{365}\right)^{n-1}$

- Conditional Probability: $P(A|B) = \frac{P(A \cap B)}{P(B)} = P(A, \text{given } B)$
- **Multiplicative law**: $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$
  - $P(A \cap B \cap C \cap D) = P(A)P(B|A)P(C|A \cap B)P(D|A \cap B \cap C)$
- Independent events: $P(A|B) = P(A)$; $P(A \cap B) = P(A)P(B)$
- **Bayes Theorem**: Find $P(A_j|B) \rightarrow$ If B occurred, has $A_j$ occurred?

List out all cases ($A_i \rightarrow$B); $P(A|B) = \frac{P(B \cap A_j) = P(A_j)P(B|A_j)}{\sum_{j=1}^{k} P(B \cap A_j) = P(A_j)P(B|A_j)}$

80% of mail are "A: {spam}"; $P(A) = 0.8$
In spam, 0.2 have "free" $P(B|A) = 0.2$
In non-spam, 0.01 "free" $P(B|\bar{A}) = 0.01$

There are only two cases resulting in B: Spam ($A$) and not Spam ($\bar{A}$)

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

P(spam|"free") = $\frac{0.8 \times 0.2}{0.8 \times 0.2 + (1 - 0.8) \times 0.01}$
$= 0.9877$

## Normal Distribution

Aim: Determine probabilities of where the datum lie in the model

Population Mean $\mu$ or $p$: $\frac{1}{N} \sum_{i=1}^{N} Y_i$

Sample Average $\bar{Y}$ or $\hat{p}$: $\frac{1}{n} \sum_{i=1}^{n} Y_i$

Pop. Variance $\sigma^2$: $\frac{1}{N} \sum_{i=1}^{N} (Y_i - \mu)^2$

S. Variance $S^2$: $\frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \mu)^2$

P. Standard Deviation: $\sigma$; S. SD: $S$

Shape depends only on $\mu$ and $\sigma^2$;

Let $Y \sim N(\mu, \sigma^2)$; $P(A, Y, B) = 0.95$

$Z$ transformation: $P\left(\frac{A-\mu}{\sigma}, Z, \frac{B-\mu}{\sigma}\right)$

$\frac{Y-\mu}{\sigma} = Z \quad P(-1.96 < Z < 1.96) = 0.95$

$Z \sim N(0, 1)$: Pivotal Quantity 1.96: Get from app.

**Central limit theorem**: Aim: Determine the minimum sample size $n$, such that the error $(\bar{Y} - \mu) < A_{Question}$ with a large probability (0.95)

If $Y_1, \ldots, Y_n$ are independent random variables that follow a normal distribution $N(\mu, \sigma^2)$, then $\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$.

E.g: IQ Score

$0.95 = P(-A < \bar{Y} - \mu < A)$

$0.95 = P\left(\frac{-A}{\sigma/\sqrt{n}} < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < \frac{A}{\sigma/\sqrt{n}}\right)$

$0.95 = P\left(\frac{-A}{\sigma/\sqrt{n}} < Z < \frac{A}{\sigma/\sqrt{n}}\right)$

$0.95 = P(-1.96 < Z < 1.96)$

$\frac{A}{\sigma/\sqrt{n}} = 1.96$

E.g: CE Support (binary $X_i$ = {1,0})

$\hat{p} = \bar{X} = \frac{X_1 + \cdots + X_n}{n}$

$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$

$0.95 = P(-A < \hat{p} - p < A)$

$\frac{A}{\sqrt{p(1-p)}/\sqrt{n}} = 1.96$

If $p$ unknown, let $p = 0.5$ (push n)

**Confidence Interval**: Have $1 - \alpha$ (Confidence Coefficient) chance that $(\hat{\theta}_L < \theta < \hat{\theta}_U)$

$= C.I. = \hat{\theta} \pm z_{\alpha/2} \times \sqrt{Var(\hat{\theta})}$

$Q: Y \sim N(\mu, 15^2), n = 25, \bar{Y} = 100$

95% CI for $\mu$:

$= 100 \pm 1.96 \times \sqrt{\frac{15^2}{25}}$

$= (94.12, 105.88)$

$Q: \hat{p} = 0.32, n = 990$

95% CI for $p$:

$= 0.32 \pm 1.96 \times \sqrt{\frac{0.32 \times (1 - 0.32)}{990}}$

$= (0.2909, 0.3491)$

## p-Value

Aim: Reject null hypothesis ($H_0$) if alternative hypothesis ($H_1$) more likely

1. Declare $H_0$: $\theta = \theta_0$ , $H_1$: $\theta > \theta_0$
2. Calculate p-value: $P(\bar{Y} > \bar{y}_{obs})$
3. Reject $H_0$ if p-value $< \alpha$ — $\alpha$: level of significance

Meaning: chance that a sample is more supportive of $H_1$ than $H_0$

Type I error: Reject $H_0$ wrongly

Type II error: Accept $H_0$ wrongly

$H_0: \mu = 100$; $H_1: \mu > 100$; $\alpha = 0.05$

$Sample: \bar{Y} = 112, n = 30, \sigma = 15$

Assume $H_0$ is true, $Y \sim N(100, 15^2)$

p-value $= P(\bar{Y} > 112 | H_0)$

$= P\left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} > \frac{T - H_0}{\sigma/\sqrt{n}}\right)$

$= P(Z > 4.3818) = 5.89 \times 10^{-5} < \alpha = 0.05$

$|4.38| > 1.645$(RR) (from back, see α = 0.05)

$\therefore H_0$ is rejected, $H_1$ is true

$H_0: p = 0.6$; $H_1: p > 0.6$; $\alpha = 0.05$

$Sample\ data: \hat{p} = \frac{X}{n} = \frac{70}{100} = 0.7$

Assume $H_0$ is true, $\hat{p} \sim N\left(0.6, \frac{0.6(1-0.6)}{100}\right)$

p-value $= P(\hat{p} > 0.7 | H_0)$

$= P\left(\frac{\hat{p} - p_0}{\sqrt{p(1-p)}/\sqrt{n}} > \frac{T - H_0}{\sqrt{H_0(1-H_0)}/\sqrt{n}}\right)$

$= P(Z > 2.04) = 0.027 < \alpha = 0.05$

$|2.04| > 1.645$ (from back, see α = 0.05)

$\therefore H_0$ is rejected, $H_1$ is true

## $\chi^2$ Test

Aim: Describe how a theoretical hypothesis fits an observation set

$X^2 = \sum_{\text{all cells}} \frac{(Observed_i - Expected_i)^2}{Expected_i}$

Reject $H_0$ if $X^2 > \chi^2_{\alpha,(d.f.)}$ (from appendix)

Otherwise, follow trend ($H_0$)

1 row: degrees of freedom = (k-1)

$H_0$: Row follows (A,B,C)$_{Question}$ classification; $H_1$: $H_0$ is not true

Given Prob. (<1)

$H_0: p_1 = \frac{1}{4}, p_2 = \frac{1}{2}, p_3 = \frac{1}{4}$

| | A | B | C | T |
|---|---|---|---|---|
| O | 32 | 76 | 36 | 144 |
| E | 36 | 72 | 36 | 144 |

$X^2 = \frac{(32-36)^2}{36} + \frac{(76-72)^2}{72} + \frac{(36-36)^2}{36} = \frac{2}{3}$

$\frac{2}{3} = X^2 < \chi^2_{0.05,(3-1)} = 5.99$

Not rejected, follow A,B,C

r×c: deg of freedom = (r-1)×(c-1)

$H_0$: Row independent of column/ No difference; $H_1$: $H_0$ is not true

Given Freq (>1)

$41.601 = X^2 > \chi^2_{0.05,4} = 9.49$

Rejected, dependent/have relationship bet. R and C

$E_{i,j} = \frac{r_i \times c_j}{n}$

$r_i = sum(row\ i)$

$n = sum(all\ cells)$