1 Statistics

Probability

$$\begin{aligned}
\Pr(A \cup B) &= \Pr(A) + \Pr(B) - \Pr(A \cap B) \\
\Pr(A \cap B) &= \Pr(A|B) \times \Pr(B) \\
\Pr(A|B) &= \frac{\Pr(A \cap B)}{\Pr(B)} \\
\Pr(A) &= \Pr(A|B) \cdot \Pr(B) + \Pr(A|B') \cdot \Pr(B')
\end{aligned}$$

Mutually exclusive: $\Pr(A \cap B) = 0$

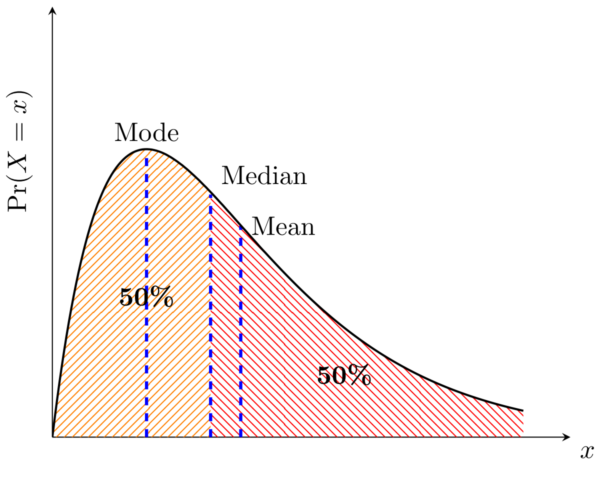
Independent events:

$$\begin{aligned}
\Pr(A \cap B) &= \Pr(A) \times \Pr(B) \\
\Pr(A|B) &= \Pr(A) \\
\Pr(B|A) &= \Pr(B)
\end{aligned}$$

Combinatorics

- Arrangements $\binom{n}{k} = \frac{n!}{(n-k)}$
- Combinations $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- Note $\binom{n}{k} = \binom{n}{k-1}$

Distributions



Mean μ

$$\begin{aligned}
E(X) &= \frac{\Sigma\,[x \cdot f(x)]}{\Sigma f} && (f = \text{absolute frequency}) \\
&= \sum_{i=1}^{n} [x_i \cdot \Pr(X = x_i)] && \text{(discrete)} \\
&= \int_{\mathbf{X}} (x \cdot f(x))\, dx
\end{aligned}$$

Mode

Value of X which has the highest probability

- Most popular value in discrete distributions
- Must exist in distribution
- Represented by local max in pdf
- Multiple modes exist when > 1 X value have equal-highest probability

Median

Value separating lower and upper half of distribution area

Continuous:

$$m = X \text{ such that } \int_{-\infty}^{m} f(x)\, dx = 0.5$$

Discrete: (not in course)

- Does not have to exist in distribution
- Add values of X smallest to largest until sum is ≥ 0.5
- If $X_1 < 0.5 < X_2$, then median is the average of X_1 and X_2
  - If $m > 0.5$, then value of X that is reached is the median of X

Variance σ^2

$$\begin{aligned}\operatorname{Var}(x) &= \sum_{i=1}^{n} p_i (x_i - \mu)^2 \\ &= \sum (x-\mu)^2 \times \Pr(X = x) \\ &= \sum x^2 \times p(x) - \mu^2 \\ &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 \\ &= E\left[(X-\mu)^2\right]\end{aligned}$$

Standard deviation σ

$$\begin{aligned}\sigma &= \operatorname{sd}(X) \\ &= \sqrt{\operatorname{Var}(X)}\end{aligned}$$

Binomial distributions

Conditions for a *binomial distribution*:

1. Two possible outcomes: **success** or **failure**
2. Pr(success) (=p) is constant across trials
3. Finite number n of independent trials

Properties of $X \sim \mathrm{Bi}(n, p)$

$$\mu(X) = np$$
$$\mathrm{Var}(X) = np(1-p)$$
$$\sigma(X) = \sqrt{np(1-p)}$$
$$\Pr(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$

On CAS

Interactive → Distribution → `binomialPdf`

x:	no. of successes
numtrial:	no. of trials
pos:	probability of success

Continuous random variables

A continuous random variable X has a pdf f such that:

1. $f(x) \geq 0 \forall x$
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$

$$E(X) = \int_{\mathbf{X}} (x \cdot f(x))\, dx$$
$$\mathrm{Var}(X) = E\left[(X - \mu)^2\right]$$

$$\Pr(X \leq c) = \int_{-\infty}^{c} f(x)\, dx$$

On CAS

Define piecewise functions:

Math3 → 

Two random variables X, Y

If X and Y are independent:

$$\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$$
$$\mathrm{Var}(aX \pm bY \pm c) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y)$$

Linear functions $X \to aX + b$

$$\begin{aligned}
\Pr(Y \le y) &= \Pr(aX + b \le y) \\
&= \Pr\left(X \le \frac{y-b}{a}\right) \\
&= \int_{-\infty}^{\frac{y-b}{a}} f(x)\, dx
\end{aligned}$$

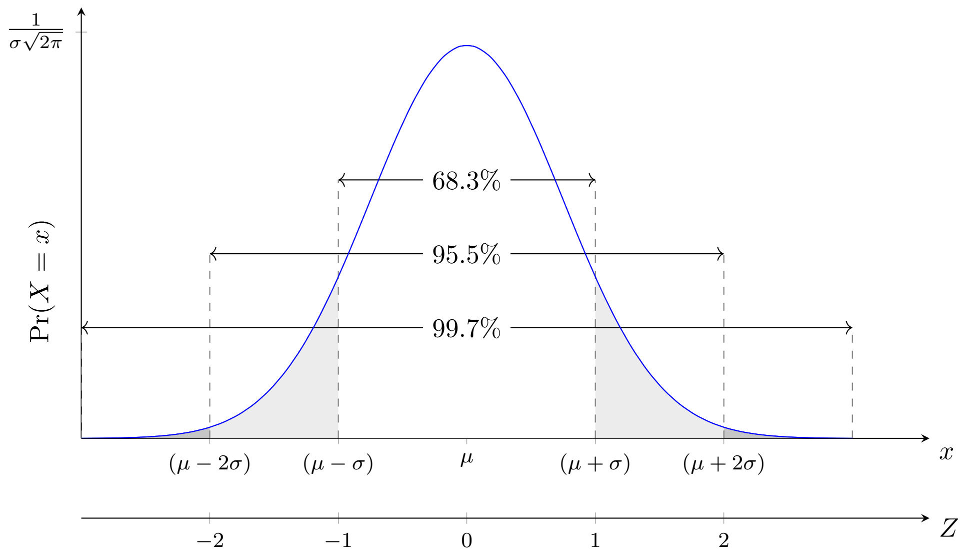
Mean: $\mathrm{E}(aX+b) = a\,\mathrm{E}(X) + b$

Variance: $\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X)$

Expectation theorems

For some non-linear function g, the expected value $E(g(X))$ is not equal to $g(E(X))$.

$$\begin{aligned}
E(X^2) &= \mathrm{Var}(X) - [E(X)]^2 \\
E(X^n) &= \Sigma x^n \cdot p(x) && \text{(non-linear)} \\
&\neq [E(X)]^n \\
E(aX \pm b) &= aE(X) \pm b && \text{(linear)} \\
E(b) &= b && (\forall b \in \mathbb{R}) \\
E(X+Y) &= E(X) + E(Y) && \text{(two variables)}
\end{aligned}$$



Sample mean

Approximation of the **population mean** determined experimentally.

$$\overline{x} = \frac{\Sigma x}{n}$$

where

n is the size of the sample (number of sample points)

x is the value of a sample point

On CAS

1. Spreadsheet
2. In cell A1:

   `mean(randNorm(sd, mean, sample size))`
3. Edit → Fill → Fill Range
4. Input range as A1:An where n is the number of samples
5. Graph → Histogram

Sample size of n

$$\overline{X} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{\sum x}{n}$$

Sample mean is distributed with mean μ and sd $\frac{\sigma}{\sqrt{n}}$ (approaches these values for increasing sample size n).

For a new distribution with mean of n trials, $\mathrm{E}(X') = \mathrm{E}(X)$, $\quad \mathrm{sd}(X') = \dfrac{\mathrm{sd}(X)}{\sqrt{n}}$

On CAS

- Spreadsheet → Catalog → `randNorm(sd, mean, n)` where `n` is the number of samples. Show histogram with Histogram key in top left
- To calculate parameters of a dataset: Calc → One-variable

Population sampling

Population proportion

$$p = \frac{n \text{ with attribute in population}}{\text{population size}}$$

Constant for a given population.

Sample proportion

$$\hat{p} = \frac{n \text{ with attribute in sample}}{\text{sample size}}$$

Varies with each sample.

Normal distributions

$$Z = \frac{X - \mu}{\sigma}$$

Normal distributions must have area (total prob.) of 1 $\implies \int_{-\infty}^{\infty} f(x)\,dx = 1$

mean = mode = median

Always express z as +ve. Express confidence interval as ordered pair.

Confidence intervals

- **Point estimate:** single-valued estimate of the population mean from the value of the sample mean $\overline{x}$
- **Interval estimate:** confidence interval for population mean μ
- $C\%$ confidence interval $\implies$ $C\%$ of samples will contain population mean μ

On CAS

Menu → Stats → Calc → Interval

Set *Type = One-Sample Z Int*

and select *Variable*

95% confidence interval

For 95% c.i. of population mean μ:

$$x \in \left(\overline{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}\right)$$

where:

$\overline{x}$ is the sample mean

σ is the population sd

n is the sample size from which $\overline{x}$ was calculated

Confidence interval of p from $\hat{p}$

$$x \in \left(\hat{p} \pm Z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Margin of error

For 95% confidence interval of μ:

$$\begin{aligned} M &= 1.96 \times \frac{\sigma}{\sqrt{n}} \\ &= \frac{1}{2} \times \text{width of c.i.} \\ \implies n &= \left(\frac{1.96\sigma}{M}\right)^2 \end{aligned}$$

Always round n up to a whole number of samples.

General case

For $C\%$ c.i. of population mean μ:

$$x \in \left(\overline{x} \pm k \frac{\sigma}{\sqrt{n}} \right)$$

where k is such that $\Pr(-k < Z < k) = \frac{C}{100}$

On CAS

Menu → Stats → Calc → Interval

Set *Type* = *One-Prop Z Int*

Input x = $\hat{p} * n$

Confidence interval for multiple trials

For a set of n confidence intervals (samples), there is 0.95^n chance that all n intervals contain the population mean μ.