# 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 $\implies \Pr(A \cup B) = 0$

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 $\mu$**

**Mean $\mu$** or **expected value** $E(X)$

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

**Mode**

Most popular value (has highest probability of all $X$ values). Multiple modes can exist if $> 1$ $X$ value have equal-highest probability. Number must exist in distribution.

**Median**

If $m > 0.5$, then value of $X$ that is reached is the median of $X$. If $m = 0.5 = 0.5$, then $m$ is halfway between this value and the next. To find $m$, add values of $X$ from smallest to alrgest until the sum reaches 0.5.

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

**Variance** $\sigma^2$

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

**Standard deviation** $\sigma$

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

## Binomial distributions

Conditions for a *binomial distribution*:

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

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

$$\begin{aligned} \mu(X) &= np \\ \text{Var}(X) &= np(1-p) \\ \sigma(X) &= \sqrt{np(1-p)} \\ \Pr(X = x) &= \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \end{aligned}$$

On CAS

Interactive → Distribution → `binomialPdf` then input

| | |
|---|---|
| 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. $\int_{-\infty}^{\infty} f(x)\,dx = 1$

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

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

## 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 $g(E(X))$.

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

### 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 $\mu$ 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

### 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 $\mu$

- $C\%$ confidence interval $\implies$ $C\%$ of samples will contain population mean $\mu$

**95% confidence interval**

For 95% c.i. of population mean $\mu$:

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

where:

$\overline{x}$ is the sample mean

$\sigma$ is the population sd

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

On CAS

Menu → Stats → Calc → Interval

Set *Type = One-Sample Z Int*

and select *Variable*

## Margin of error

For 95% confidence interval of $\mu$:

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

Always round $n$ up to a whole number of samples.

## General case

For $C\%$ c.i. of population mean $\mu$:

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

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

On CAS

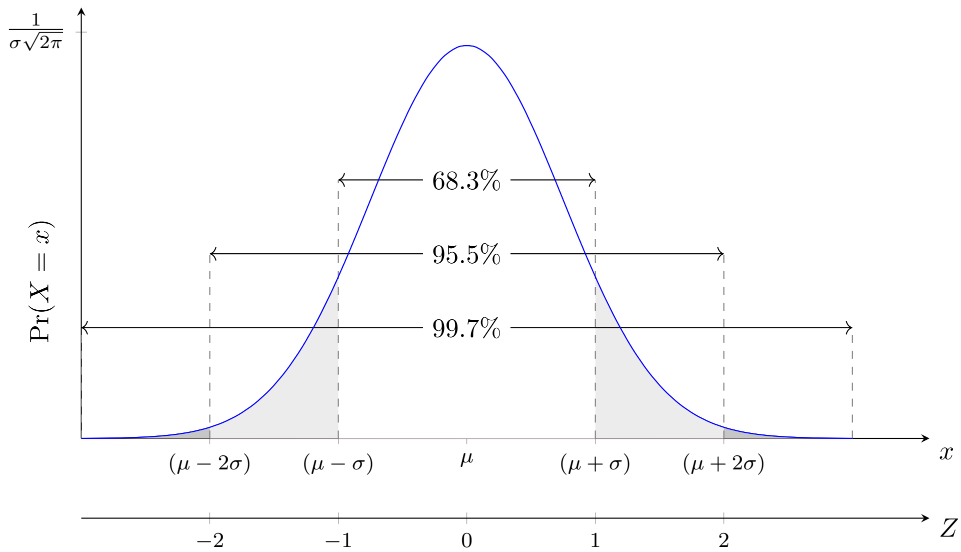
Menu → Stats → Calc → Interval

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 $\mu$.


$\frac{1}{\sigma\sqrt{2\pi}}$
$\Pr(X = x)$
68.3%
95.5%
99.7%
$(\mu - 2\sigma)$
$(\mu - \sigma)$
$\mu$
$(\mu + \sigma)$
$(\mu + 2\sigma)$
$x$
$-2$
$-1$
0
1
2
$Z$