

# 1 Statistics

## Continuous random variables

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

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

$$E(X) = \int_{\mathbf{X}} (x \cdot f(x)) dx$$

$$\text{Var}(X) = E[(X - \mu)^2]$$

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

## Two random variables $X, Y$

If  $X$  and  $Y$  are independent:

$$E(aX + bY) = aE(X) + bE(Y)$$

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

## Linear functions $X \rightarrow aX + b$

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

$$\textbf{Mean:} \quad E(aX + b) = aE(X) + b$$

$$\textbf{Variance:} \quad \text{Var}(aX + b) = a^2 \text{Var}(X)$$

## Expectation theorems

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

$$E(X^2) = \text{Var}(X) + [E(X)]^2$$

$$E(X^n) = \sum x^n \cdot p(x) \quad (\text{non-linear})$$

$$\neq [E(X)]^n$$

$$E(aX \pm b) = aE(X) \pm b \quad (\text{linear})$$

$$E(b) = b \quad (\forall b \in \mathbb{R})$$

$$E(X + Y) = E(X) + E(Y) \quad (\text{two variables})$$

## Sample mean

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

$$\bar{x} = \frac{\sum x}{n}$$

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 $n$

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

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,  $E(X') = E(X)$ ,  $\text{sd}(X') = \frac{\text{sd}(X)}{\sqrt{n}}$

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

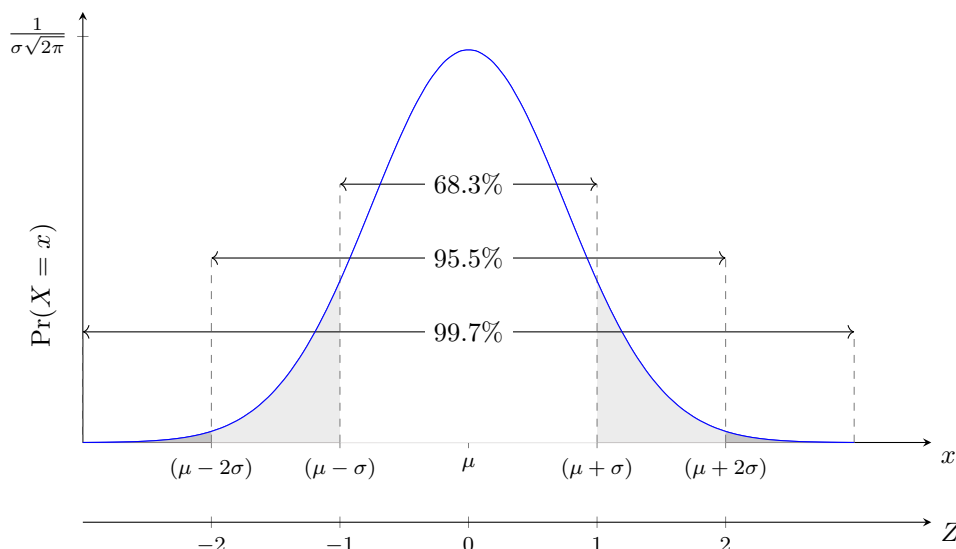
## Normal distributions

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

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.**



## Central limit theorem

If  $X$  is randomly distributed with mean  $\mu$  and sd  $\sigma$ , then with an adequate sample size  $n$  the distribution of the sample mean  $\bar{X}$  is approximately normal with mean  $E(\bar{X})$  and  $\text{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ .

## Confidence intervals

- **Point estimate:** single-valued estimate of the population mean from the value of the sample mean  $\bar{x}$
- **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  $\mu$ :

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

where:

$\bar{x}$  is the sample mean

$\sigma$  is the population sd

$n$  is the sample size from which  $\bar{x}$  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  $\mu$ :

$$M = 1.96 \times \frac{\sigma}{\sqrt{n}}$$

$$\implies n = \left( \frac{1.96\sigma}{M} \right)^2$$

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

## General case

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

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

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

## 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  $\mu$ .

## 2 Hypothesis testing

Note hypotheses are always expressed in terms of population parameters

### Null hypothesis $H_0$

Sample drawn from population has same mean as control population, and any difference can be explained by sample variations.

### Alternative hypothesis $H_1$

Amount of variation from control is significant, despite standard sample variations.

***p*-value**

$$\begin{aligned}
 p\text{-value} &= \Pr(\bar{X} \leq \mu(\mathbf{H}_1) \mid \mu = \mu(\mathbf{H}_0)) \\
 &= \Pr\left(Z \leq \frac{(\mu(\mathbf{H}_1) - \mu(\mathbf{H}_0)) \cdot \sqrt{n}}{\text{sd}(X)}\right)
 \end{aligned}$$

then use `normCdf` with `std. norm.`

Probability of observing a value of the sample statistic as significant as the one observed, assuming null hypothesis is true.

<b><i>p</i></b>	<b>Conclusion</b>
> 0.05	insufficient evidence against $\mathbf{H}_0$
< 0.05 (5%)	good evidence against $\mathbf{H}_0$
< 0.01 (1%)	strong evidence against $\mathbf{H}_0$
< 0.001 (0.1%)	very strong evidence against $\mathbf{H}_0$

**Significance level  $\alpha$** 

The condition for rejecting the null hypothesis.

If  $p < \alpha$ , null hypothesis is **rejected**

If  $p > \alpha$ , null hypothesis is **accepted**

***z*-test**

Hypothesis test for a mean of a sample drawn from a normally distributed population with a known standard deviation.

On CAS

Menu  $\rightarrow$  Statistics  $\rightarrow$  Calc  $\rightarrow$  Test.

Select *One-Sample Z-Test* and *Variable*, then input:

$\mu$  cond: same operator as  $\mathbf{H}_1$

$\mu_0$ : expected sample mean (null hypothesis)

$\sigma$ : standard deviation (null hypothesis)

$\bar{x}$ : sample mean

$n$ : sample size

**One-tail and two-tail tests**

$$p\text{-value (two-tail)} = 2 \times p\text{-value (one-tail)}$$

**One tail**

- $\mu$  has changed in one direction
- State " $\mathbf{H}_1 : \mu \leq$  known population mean"

**Two tail**

- Direction of  $\Delta\mu$  is ambiguous
- State “ $\mathbf{H}_1 : \mu \neq$  known population mean”

$$p\text{-value} = \Pr(|\bar{X} - \mu| \geq |\bar{x}_0 - \mu|)$$

$$= \left( |Z| \geq \left| \frac{\bar{x}_0 - \mu}{\sigma \div \sqrt{n}} \right| \right)$$

where

$\mu$  is the population mean under  $\mathbf{H}_0$

$\bar{x}_0$  is the observed sample mean

$\sigma$  is the population s.d.

$n$  is the sample size

**Modulus notation for two tail**

$\Pr(|\bar{X} - \mu| \geq a) \implies$  “the probability that the distance between  $\bar{\mu}$  and  $\mu$  is  $\geq a$ ”

**Inverse normal**

On CAS

```
invNormCdf("L",  $\alpha$ ,  $\frac{\sigma}{n^\alpha}$ ,  $\mu$ )
```

**Errors**

**Type I error**  $\mathbf{H}_0$  is rejected when it is **true**

**Type II error**  $\mathbf{H}_0$  is **not** rejected when it is **false**

	Actual result	
	$\mathbf{H}_0$ true	$\mathbf{H}_0$ false
<b>z-test</b>		
Reject $\mathbf{H}_0$	Type I error	Correct
Do not reject $\mathbf{H}_0$	Correct	Type II error