Statistics

1 Linear combinations of random variables

Continuous random variables

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

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

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

Linear functions $X \to aX + b$

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

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

Linear combination of two random variables

Mean:	$\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$	
Variance:	$\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y)$	(if X and Y are independent)

2 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) and 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

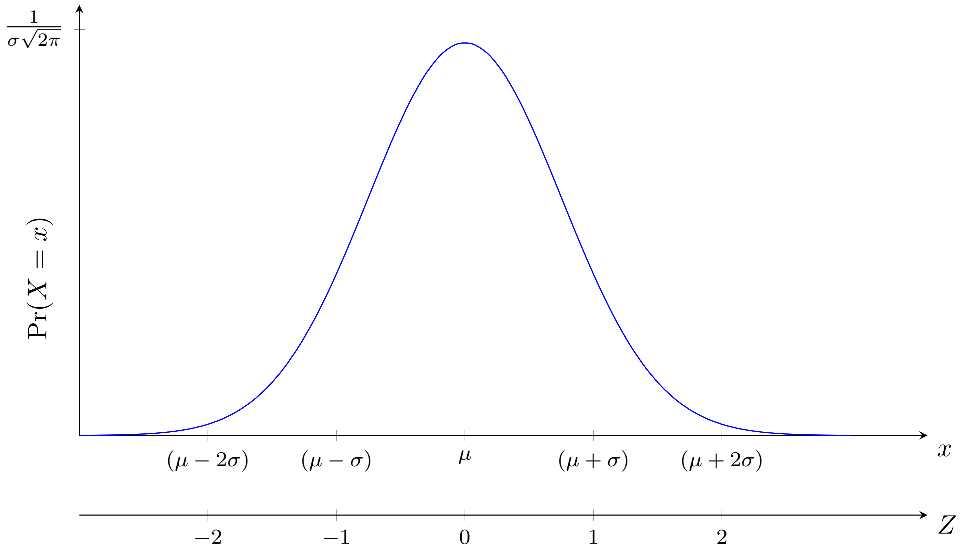
- Spreadsheet → Catalog → `randNorm(sd, mean, n)` where `n` is the number of samples. Show histogram with Histogram key in top left
- To calculate parameters of a dataset: Calc → One-variable

3 Normal distributions

mean = mode = median

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

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



4 Central limit theorem

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

5 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

5.1 95% confidence interval

The 95% confidence interval for a population mean μ is given by

$$\overline{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

where:
$\overline{x}$ is the sample mean
σ is the population sd
n is the sample size from which $\overline{x}$ was calculated
Always express z as +ve. Express confidence *interval* as ordered pair.
On CAS
Menu → Stats → Calc → Interval
Set Type = One-Sample Z Int, Variable

Interpretation of confidence intervals

95% confidence interval $\implies$ 95% of samples will contain population mean μ.

Margin of error

For 95% confidence interval for μ, margin of error M is:

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

General case

A confidence interval of $C\%$ for a mean μ s given by

$$x \in \left(\overline{x} \pm k\frac{\sigma}{\sqrt{n}}\right) \qquad \text{where } k \text{ is such that } \Pr(-k < Z < k) = \frac{C}{100}$$

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 μ.

6 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

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

Distribution of sample mean

If $X \sim \text{N}(\mu, \sigma)$, then distribution of sample mean $\overline{X}$ is also normal with $\overline{X} \sim \text{N}(\mu, \frac{\sigma}{\sqrt{n}}$.

Statistical significance

Significance level is denoted by α.
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 → Statistics → Calc → Test.
Select *One-Sample Z-Test* and *Variable*, then input:

- μ condition - same operator as H_1
- μ_0 - expected sample mean (null hypothesis)
- σ - standard deviation (null hypothesis)
- $\overline{x}$ - sample mean
- n - sample size