Statistics

Consider a sequence of N quantitative data numbers:

$$x_1, x_2, x_3, \ldots, x_n$$

Summation / Product notation

$$\sum_{k=1}^{N} x_k = x_1 + x_2 + x_3 + \cdots + x_N$$

$$\prod_{k=1}^{N} x_k = x_1 x_2 x_3 \cdots x_N .$$

Statistical characterizations of data

(1) The average $\bar{x}$:

$$\boxed{\bar{x} = \frac{1}{N} \sum_{k=1}^{N} x_k}$$

Notation: $E(x) \equiv \bar{x}$

② Standard deviation → Estimates how much the data tends to deviate from the statistical average

$$\sigma_x = \sqrt{\frac{1}{N} \sum_{a=1}^{N} (x_a - \bar{x})^2}$$

$$s_x = \sqrt{\frac{1}{N-1} \sum_{a=1}^{N} (x_a - \bar{x})^2}$$

Here
σ_x = population standard deviation
(used when the data is complete)
s_x = sample standard deviation
(used when the data is a SAMPLE of the complete dataset)

③ Variance → Is defined as the square of population standard deviation

$$Var(x) = \frac{1}{N} \sum_{a=1}^{N} (x_a - E(x))^2$$

$$= E((x - E(x))^2)$$

Properties of $E(X)$ and $Var(X)$

Let $X = x_1, x_2, x_3, \dots, x_N$
$Y = y_1, y_2, y_3, \dots, y_N$.

1) $\forall a \in \mathbb{R} : \forall b \in \mathbb{R} : E(aX+b) = aE(X)+b$
2) $E(X+Y) = E(X) + E(Y)$
3) $\forall a \in \mathbb{R} : E(aX) = aE(X)$.

For the variance we have:

1) $\forall a \in \mathbb{R} : \forall b \in \mathbb{R} : Var(aX+b) = a^2 Var(X)$
2) $\forall a \in \mathbb{R} : Var(aX) = a^2 Var(X)$
3) $Var(X) = E(X^2) - [E(X)]^2$.

Proof of (3)

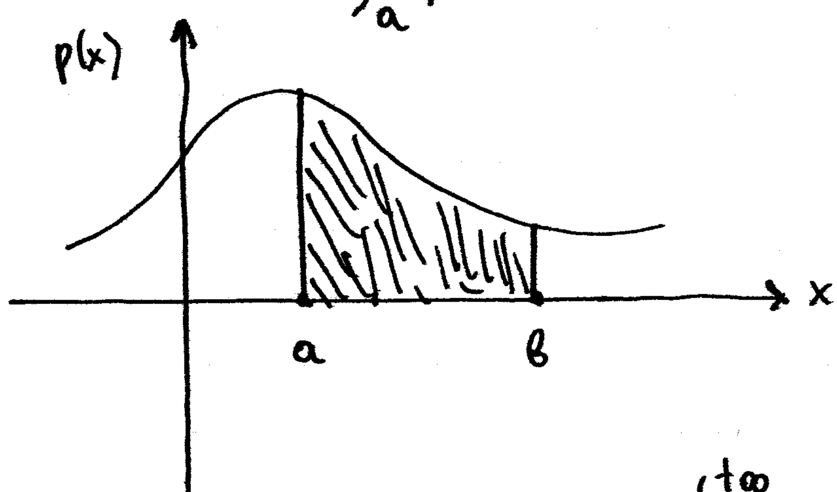
$$
\begin{aligned}
Var(X) &= E((X-E(X))^2) = \\
&= E(X^2 - 2XE(X) + (E(X))^2) \\
&= E(X^2) - 2E(XE(X)) + E((E(X))^2) \\
&= E(X^2) - 2E(X)E(X) + (E(X))^2 \\
&= E(X^2) - 2[E(X)]^2 + [E(X)]^2 \\
&= E(X^2) - [E(X)]^2.
\end{aligned}
$$

Normal random variables

Consider a random experiment with sample space $\Omega = \mathbb{R}$.

- The outcome of the experiment can be any real number $x \in \mathbb{R}$.
- We associate with the random experiment a <u>probability density function</u> $p(x)$ such that the probability $P(a \leq x \leq b)$ that x will satisfy $a \leq x \leq b$ is given by:

$$P(a \leq x \leq b) = \int_a^b p(x)\,dx$$

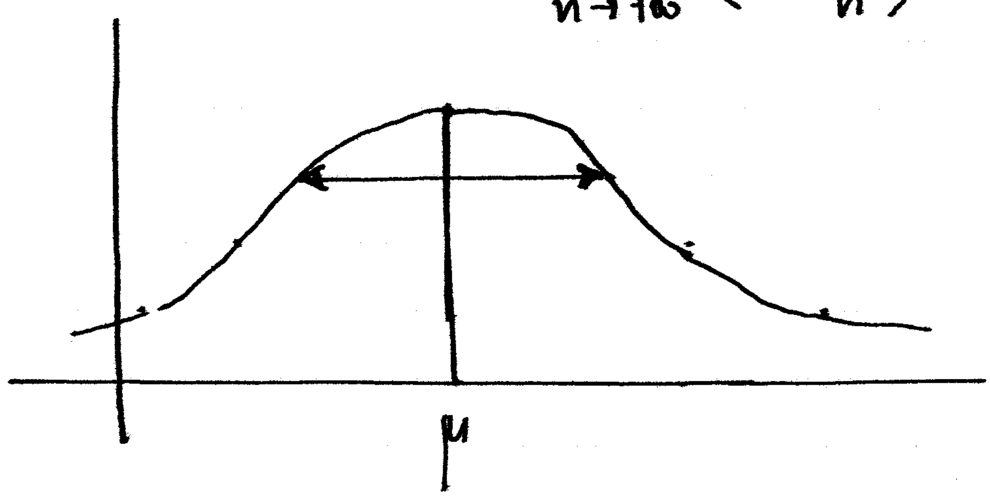


Obviously we expect that $\int_{-\infty}^{+\infty} p(x)\,dx = 1$.

Def : We say that a random variable X on sample space $\Omega = \mathbb{R}$ is a normal variable with mean μ and standard deviation σ if it has probability distribution given by

$$\boxed{\begin{aligned} p(x) &= \frac{1}{\sigma}\,\varphi\left(\frac{x-\mu}{\sigma}\right) \\ \varphi(x) &= \frac{1}{\sqrt{2\pi}}\exp(-x^2/2) \end{aligned}}$$

Here $\exp(x) = e^x = \lim_{n \to +\infty} \left(1 + \frac{x}{n}\right)^n$.



notation : If x is a normal random variable with mean μ and standard deviation σ we write

$$x \sim N(\mu, \sigma^2)$$

Remark : Assuming that $x \sim N(\mu, \sigma^2)$ and

$x_1, x_2, x_3, \dots, x_n$

is the outcome of a repeated random experiment, then

a) The mean of x_k approaches μ

b) The variance of x_k approaches σ^2

when $n \to +\infty$.

Properties of normal variables

1) $x \sim N(\mu, \sigma^2) \Rightarrow \forall a, b \in \mathbb{R} : ax + b \sim N(a\mu + b, (a\sigma)^2)$

2) If x, y are <u>independent</u> random variables then

$$\boxed{\left.\begin{array}{l} x \sim N(\mu_1, \sigma_1^2) \\ y \sim N(\mu_2, \sigma_2^2) \end{array}\right\} \Rightarrow x + y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)}$$

z-scores and their interpretation

- Let x be a random variable with sample

  $x_1, x_2, x_3, \dots, x_N$.

  If $\mu = E(x)$ = mean / average of x

  $\sigma = \sqrt{Var(x)}$ = standard deviation of x

  then the <u>z-score</u> of x_k is defined as

$$\boxed{z_k = \frac{(x_k - \mu)}{\sigma}}$$

▸ <u>Interpretation</u>: The z-score measures objectively how much x_k deviates from the mean μ.

▸ Assume that x is a normal variable. The probability that one experiment will give outcome x_k with z-score $-a < z_k < a$ is:

$$P(-a < z_k < a) = \int_{-a}^{a} \varphi(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{-a}^{a} \exp(-x^2/2)\, dx$$

This gives:

z-score	probability (normal variable)
$-1 \leq z \leq 1$	$\approx 68\%$
$-2 \leq z \leq 2$	$\approx 95\%$
$-3 \leq z \leq 3$	$\approx 99.7\%$

- If we do NOT assume that x is a normal variable and we do not know its probability density function $p(x)$ then,
the best we can say is that

$$\boxed{P(-a \leq z \leq a) \geq 1 - 1/a^2}$$ (Chebyshev inequality)

This gives:

z-score	Chebyshev probability (lower bound)
$-1 \leq z \leq 1$	50%
$-2 \leq z \leq 2$	75%
$-3 \leq z \leq 3$	89%
$-4 \leq z \leq 4$	94%
$-5 \leq z \leq 5$	96%

▶ z-scores can be used for <u>standardized</u> comparison of two data points from two distinct datasets.

▶ <u>example</u>

Lt. Worf has scored 80 on an exam where the mean was $\mu_1 = 60$ and $\sigma_1 = 10$.
Cmd. Data has scored 70 on an exam where the mean was $\mu_2 = 50$ and $\sigma_2 = 20$.
Who's done "better"?

<u>Solution</u>

For Lt. Worf:

$$z_1 = \frac{80 - \mu_1}{\sigma_1} = \frac{80 - 60}{10} = 2$$

For Cmd. Data:

$$z_2 = \frac{70 - \mu_2}{\sigma_2} = \frac{70 - 50}{20} = 2$$

Thus they performed equally

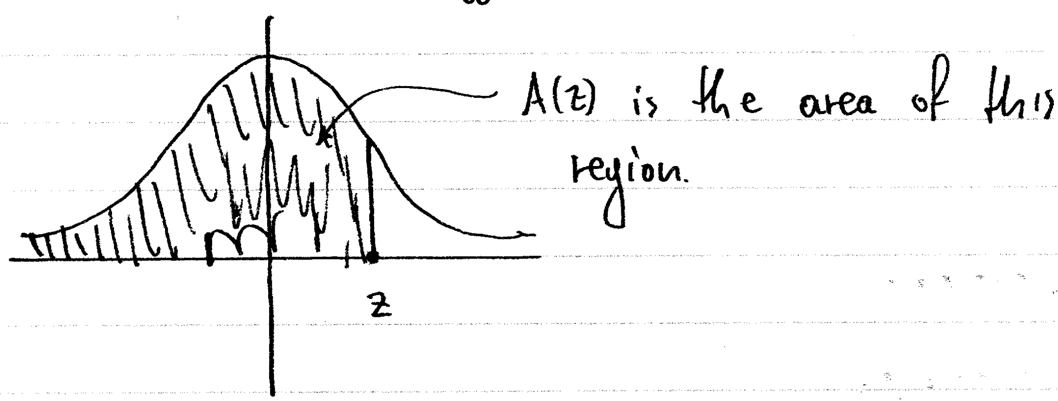
▶ Note: Bigger z means better performance.

Using the z-table

Let $x \sim N(\mu, \sigma^2)$ be a normal random variable.
Want: probability that one experiment will yield an outcome x within a certain range of values.

- The z-table evaluates the function

$$A(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx$$

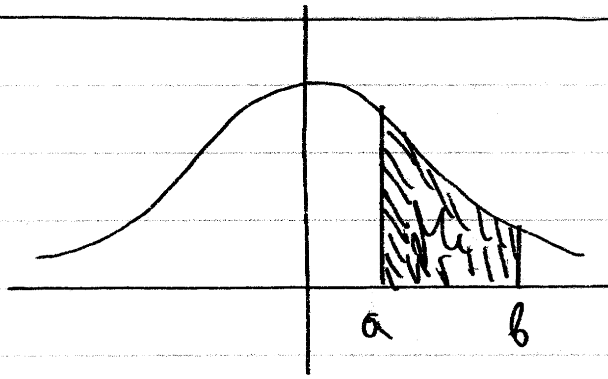


Note that $A(0) = 1/2$

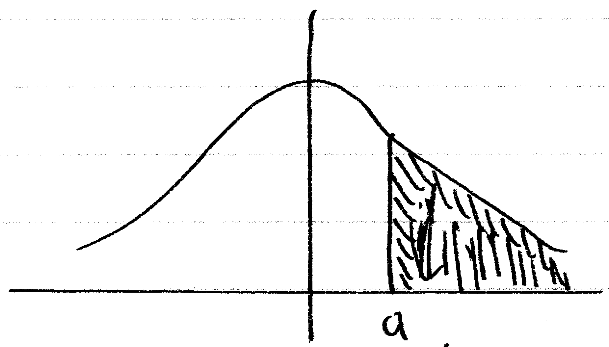
$\lim_{x \to +\infty} A(z) = 1, \quad \lim_{z \to -\infty} A(z) = 0.$

- The z-table can be used to calculate
  a) $P(a < x < b)$ = probability that $a < x < b$
  b) $P(a < x)$ = probability that $a < x$
  c) $P(x < b)$ = probability that $x < b$.

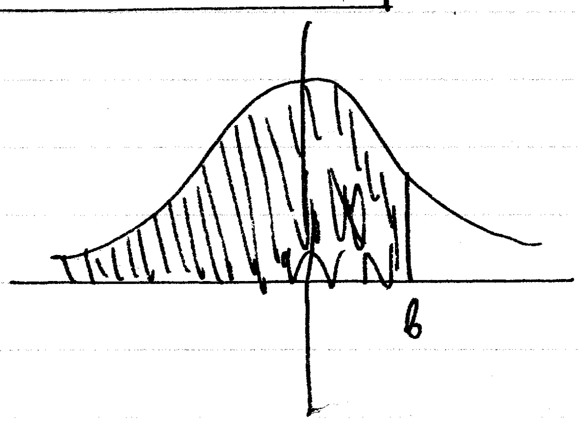
a) $\boxed{P(a<x<b) = A\left(\frac{b-\mu}{\sigma}\right) - A\left(\frac{a-\mu}{\sigma}\right)}$



b) $\boxed{P(a<x) = 1 - A\left(\frac{a-\mu}{\sigma}\right)}$



c) $\boxed{P(x<b) = A\left(\frac{b-\mu}{\sigma}\right)}$



examples

1) The daily production of bananas is normal random variable with $\mu = 1200$ and $\sigma = 100$. Find

a) Probability to produce less than 1000 bananas

Solution

$$P(x < 1000) = A\left(\frac{1000-\mu}{\sigma}\right) = A\left(\frac{1000-1200}{100}\right) =$$

$$= A(-2) = 0.0228 = 2.28\%$$

b) Probability to produce more than 1300 bananas.

Solution

$$P(x > 1300) = 1 - A\left(\frac{1300-\mu}{\sigma}\right) = 1 - A\left(\frac{1300-1200}{100}\right)$$

$$= 1 - A(1) = 1 - 0.8413 = 0.1587 = 15.87\%$$

c) Probability to produce between 1150 and 1450 bananas.

Solution

$$P(1150 < x < 1450) = A\left(\frac{1450-\mu}{\sigma}\right) - A\left(\frac{1150-\mu}{\sigma}\right)$$

$$= A\left(\frac{1450-1200}{100}\right) - A\left(\frac{1150-1200}{100}\right)$$

$$= A(2.5) - A(-0.5) = 0.9938 - 0.3085 =$$

$$= 0.6853 = 68.53\%$$