

Measures of Central Tendency and Dispersion [ST&D p. 16-27]

The individual values of a **population** are denoted Y_i , $i = 1, \dots, N$, where N is the size of the population.

The individual values of a **sample** are also denoted Y_i , but in this case $i = 1, \dots, r$, where r is the size of the sample.

Mean

- Population mean $\mu = \frac{\sum_{i=1}^N Y_i}{N}$ Sample mean $\bar{Y} = \frac{\sum_{i=1}^r Y_i}{r}$

Variance

- Population variance $\sigma^2 = \frac{\sum_{i=1}^N (Y_i - \mu)^2}{N}$
- Sample variance $s^2 = \frac{\sum_{i=1}^r (Y_i - \bar{Y})^2}{r - 1}$

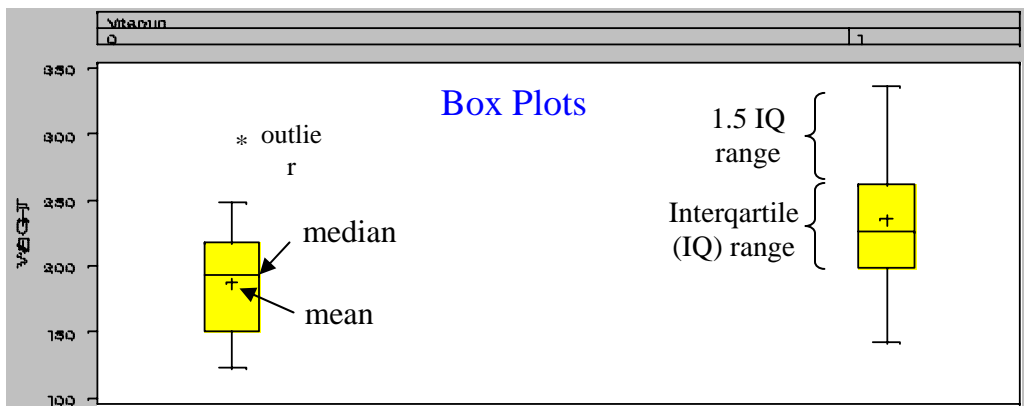
The quantities $(Y_i - \bar{Y})$ are called the *deviations*.

- Sample variance of the mean $s_{\bar{Y}}^2 = \frac{s^2}{n}$

The standard deviation of a mean is often called **standard error** $s_{\bar{Y}} = \frac{s}{\sqrt{n}}$

The SE determines the length of confidence intervals and power of tests

- Coefficient of variation** $CV = s / \bar{Y}$



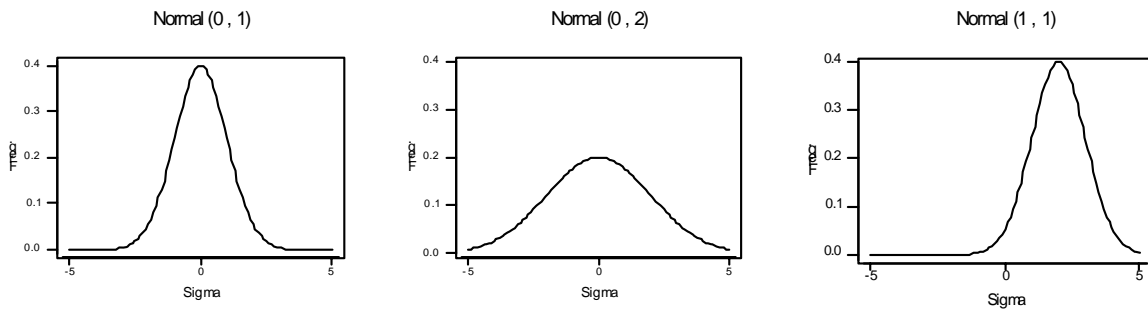
Review ST&D p. 58 Estimation and inference, p53: 3.8 Distribution of means

The normal distribution (~N)

Formally, the normal probability density function can be represented by the expression

$$Z = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}[(Y-\mu)/\sigma]^2}$$

Z indicates the density of the items. The parametric mean μ and the parametric standard deviation σ , determine the location and shape of the distribution.



The normal curve is bell-shaped and symmetrical around the mean. To convert any ~N into a standard N curve:

Standard N curve $\mu=0, \sigma=1$ \rightarrow $\frac{(Y_i-\mu)}{\sigma}$ where $-\mu$ centers to 0
 σ / σ puts variation in units of σ

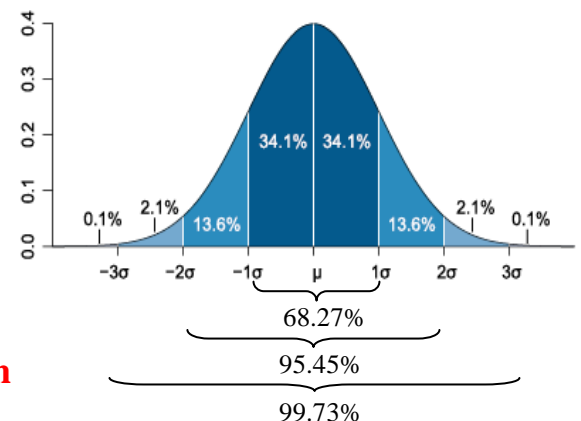
The following % of items in a N distribution lie within the indicated limits:

- $\mu \pm \sigma$ contains 68.27% of the items
- $\mu \pm 2\sigma$ contains 95.45% of the items
- $\mu \pm 3\sigma$ contains 99.73% of the items

Conversely:

- 50% of the items fall between $\mu \pm 0.674\sigma$
- 95% of the items fall between $\mu \pm 1.960\sigma$
- 99% of the items fall between $\mu \pm 2.576\sigma$

Why is it so important? **The central limit theorem**



As sample size increases, the means of samples drawn from a population of any distribution will approach the normal distribution with mean μ and variance

$$\sigma^2/r.$$

The distribution of means

Parent population

Y distribution $N(\mu, \sigma^2)$

Based on the central limit theorem

As sample size increases, the means of samples drawn from a population of **any** distribution will approach the normal distribution with mean μ and variance σ^2/r .

Derived population of sample means

\bar{Y} with distribution $N(\mu, \sigma^2/r)$

If we know σ^2 we can answer the frequent question: What is the probability that the sample mean will exceed a given value?

Example:

Variety A of oranges have $\mu=10, \sigma^2=4$

What is the probability that the mean of a random sample of $n=16$ will exceed 11?

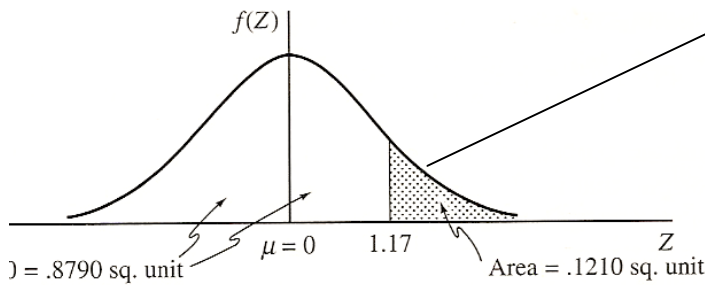
$$z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}} = \frac{11 - 10}{\sqrt{\frac{4}{16}}} = 2 \quad P(\bar{Y} \geq 11) = P(Z \geq 2)$$

Look in [Table A.4](#) $P(\bar{Y} \geq 11) = P(Z \geq 2) = 0.0228$

And this is expected because 11 is **2 standard errors** from the mean

$$s_{\bar{Y}} = \frac{s}{\sqrt{n}} = 2 / 4 = 0.5$$

Use of the normal distribution table



From Table

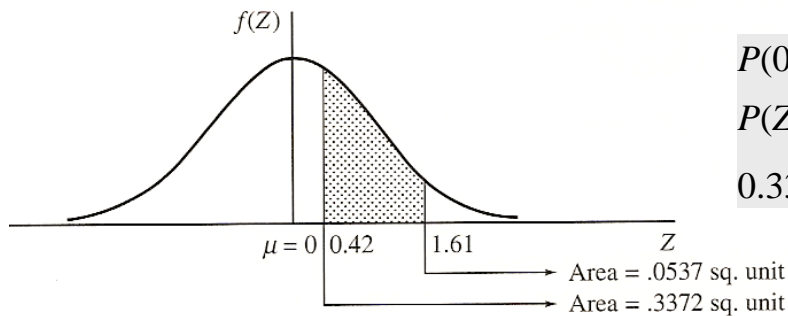
$$P(Z \geq 1.17) = 0.121 \text{ (pb inside Table)}$$

If asked

$$P(Z \leq 1.17) = 1 - P(Z \geq 1.17) = 0.879$$

or

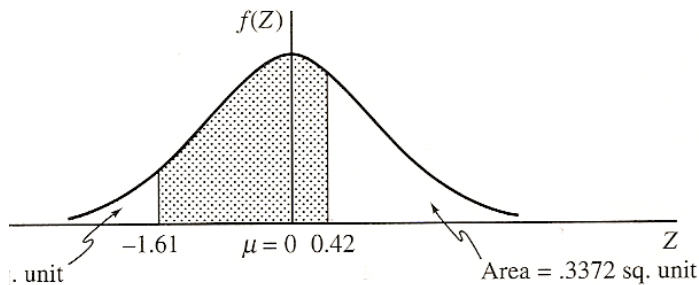
$$P(Z \geq -1.17) = 1 - P(Z \geq 1.17) = 0.879$$



$$P(0.42 \leq Z \leq 1.61) =$$

$$P(Z \geq 0.42) - P(Z \geq 1.61) =$$

$$0.3372 - 0.0537 = 0.2835$$



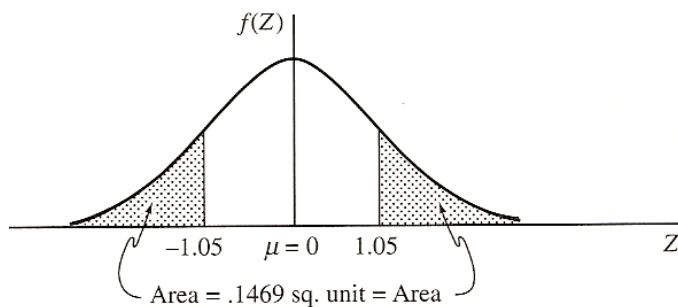
$$P(-1.61 \leq Z \leq 0.42) =$$

$$P(Z \geq -1.61) - P(Z \geq 0.42) =$$

$$1 - P(Z \geq 1.61) - P(Z \geq 0.42) =$$

$$[1 - 0.0537] - 0.3372 =$$

$$0.9463 - 0.3372 = 0.6091$$

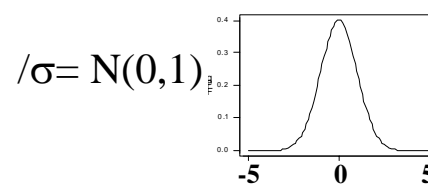
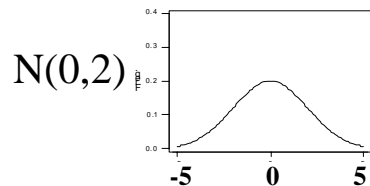
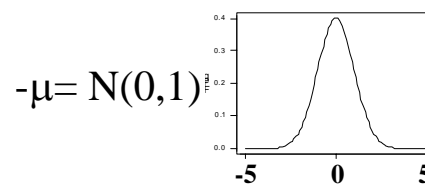
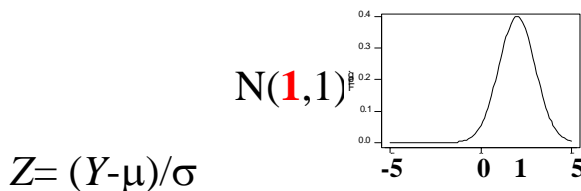


$$P(|Z| \geq 1.05) =$$

$$2 * P(Z \geq 1.05) =$$

$$2 * 0.1469 = 0.2938$$

Location and Scale transformation (when $\mu \neq 0$ and/or $\sigma \neq 1$)



Normal probability plot (Q-Q plot) ST&D p. 566

14 malt extract values: 77.7, 76.0, 76.9, 74.6, 74.7, 76.5, 74.2, 75.4, 76.0, 76.0, 73.9, 77.4, 76.6, 77.3 (ST&D p. 30, Lab1). N=14 \Rightarrow

Divide $\sim N$ in 14 intervals = area.

Normal line: slope=s=1.227, intercept= \bar{Y} =75.943. $y = a + bx$

