The sample median

Assume that the observations are ranked from the smallest ~~to the smallest~~ to the largest as:

$$x_{(1)},\ x_{(2)},\ \dots,\ x_{(n)}$$

The sample median, $\tilde{x}$, is then the observation in the middle.

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right), & \text{if } n \text{ is even} \end{cases}$$

The sample median is more robust against outliers than the average. It represents the value for which 50% of the observations are smaller and is also called the 2. sample quartile, Q_2

The first quartile, Q_1, represents the value for which 25% of the observations are smaller.

Similarly Q_3 (3. quartile) represents the value for which 75% of the observations are smaller.

The way MINITAB finds Q_1 and Q_3 is:

Q_1 is the value in position $\frac{n+1}{4}$

Q_3 is the value in position $3\left(\frac{n+1}{4}\right)$

In the case of 10 observations

$\frac{n+1}{4} = 2,75$, $3\left(\frac{n+1}{4}\right) = 8,25$

$\Rightarrow Q_1 = \frac{1}{4} X_{(2)} + \frac{3}{4} X_{(3)}$ and $Q_3 = \frac{3}{4} X_{(8)} + \frac{1}{4} X_{(9)}$

Box-plot

Compute Q_2 = sample median

Compute $h = Q_3 - Q_1$ = the height of the box

Observations further away than $\frac{3}{2}h$ from the top or the bottom of the box are called outliers.

Those further away than $3h$ are called strong outliers

A line is drawn from the top and the bottom of the box to the largest and the smallest observation that is not an outlier.

Normal-plot

Let $X \sim N(\mu, \sigma^2)$

$$F_X(x) = P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

$N(0,1)$

$$\Longrightarrow \Phi^{-1}(F_X(x)) = \frac{x-\mu}{\sigma}$$

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution. Ordered by size the values (realizations) can be written as

$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

Then $F_X(x_{(i)}) \approx \frac{\# x_j \leq x_{(i)}}{n} = \frac{i}{n} \approx \begin{cases} \frac{i-\frac{1}{2}}{n} \\ \frac{i-3/8}{n+1/4} \end{cases} = f_i$

A plot of $\Phi^{-1}(f_i)$ against $x_{(i)}$ should approximately become a straight line

With method A Inference from the use of method A

$X_1, \ldots, X_{10} \sim N(\mu_A, \sigma_A^2)$, μ_A and σ_A is unknown.

Is the expected mean less than 85?

$H_0: \mu_A = 85$ $H_1: \mu_A < 85$

The random sample approach

Testobservator $\frac{\bar{X} - 85}{\frac{\sigma_A}{\sqrt{n}}} \sim N(0,1)$, but since σ_A is unknown

where $n = 10$.

we must use $T = \frac{\bar{X} - 85}{\frac{S_A}{\sqrt{n}}} \sim t_9$ Chapter 9.4, 9.5; 10.4

Here $S_A = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(X_i - \bar{X})^2}$

$t_{obs} = \frac{84.24 - 85}{\frac{2.902}{\sqrt{10}}} = -0.83$. $P(T \leq -0.83 \mid H_0) = 0.214$

No particular reason to reject H_0.

Want to find out if $\mu_B > \mu_A$

$H_0: \mu_B = \mu_A$ $\quad$ $H_1: \mu_B > \mu_A$ $\quad$ Chapter (10)

<u>The random sample approach</u>

If. $X_1, \ldots, X_{10} \sim N(\mu_A, \sigma_A^2)$ and independent

and $Y_1, \ldots, Y_{10} \sim N(\mu_B, \sigma_B^2)$ and independent and

independent of $X_1, \ldots, X_{10}$, then

$$\frac{\bar{Y} - \bar{X} - (\mu_B - \mu_A)}{\sqrt{\frac{\sigma_B^2}{10} + \frac{\sigma_A^2}{10}}} \sim N(0,1)$$

Since, $Var(\bar{Y}) = \frac{\sigma_B^2}{10}$ and $Var(\bar{X}) = \frac{\sigma_A}{10}$ $\quad$ Chapter (8.4, 9.8, 10.5)

But σ_A^2 and σ_B^2 are unknown

$S_A^2 = \frac{1}{9} \sum_{i=1}^{10} (X_i - \bar{X})^2$, $\quad S_B^2 = \frac{1}{9} \sum_{i=1}^{10} (Y_i - \bar{Y})^2$ $\quad$ Chapter (8)

If. $\sigma_A^2 = \sigma_B^2 = \sigma^2$ we have

$$S_p^2 = \frac{\sum_{i=1}^{10} (X_i - \bar{X})^2 + \sum_{i=1}^{10} (Y_i - \bar{Y})^2}{10 + 10 - 2}$$ Kap (9.8, 10.5)

and $T = \frac{\bar{Y} - \bar{X} - (\mu_B - \mu_A)}{S\sqrt{\frac{1}{10} + \frac{1}{10}}}$ is t_{18}

Under H_0: $\mu_B - \mu_A = 0$ and $t_{obs} = \frac{85.54 - 84.24}{3.3\sqrt{\frac{1}{10} + \frac{1}{10}}} = 0.88$

$P(T \geq 0.88 \mid H_0) = 0.195$

No particular reason to claim that $\mu_B > \mu_A$

Is $\sigma_A^2 = \sigma_B^2$?

If $\sigma_A^2 \neq \sigma_B^2$ we have

$$T = \frac{\bar{Y} - \bar{X} - (\mu_B - \mu_A)}{\sqrt{\frac{S_A^2}{10} + \frac{S_B^2}{10}}} \approx \text{ t-distributed with}$$

$$V = \frac{\left(\frac{S_A^2}{10} + \frac{S_B^2}{10}\right)^2}{\left(\frac{S_A^2}{10}\right)^2/9 + \left(\frac{S_B^2}{10}\right)^2/9}$$

$$t_{obs} = \frac{85.54 - 84.24}{\sqrt{\frac{(2.9)^2}{10} + \frac{(3.65)^2}{10}}} = 0.88$$

$V = 17 \Rightarrow P(T \geq t_{obs} | H_0) = 0.195$

An alternative approach

Use of historical data

For method A we have 210 data : $X_1, \ldots, X_{210}$

1: A: $X_1, \ldots, X_{10}$ B: $X_{11}, \ldots, X_{20}$

2: A: $X_2, \ldots, X_{11}$ B: $X_{12}, \ldots, X_{21}$

⋮

191: A: $X_{191}, \ldots, X_{200}$ B: $X_{201}, \ldots, X_{210}$

Only 9 out of 191 differences are greater than 1.3

$P(\bar{X}_B - \bar{X}_A > 1.3) = \frac{9}{191} = 0.047 < 0.05$ and may indicate that Method B is better than Method A