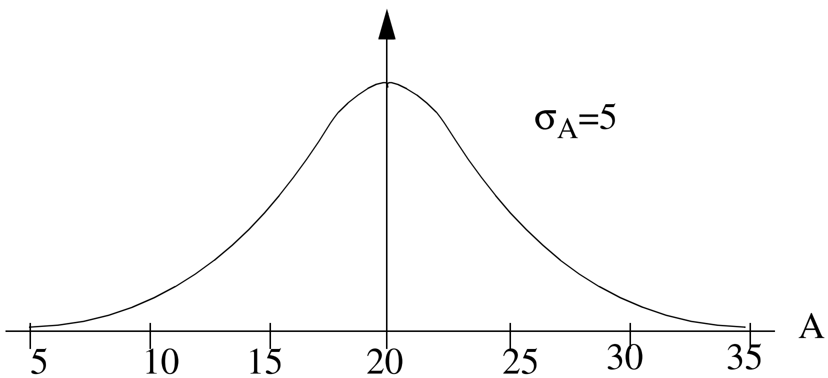
## 2.0 EXPERIMENTS COMPARING TWO VALUES
## (updated Spring 2005)

A new car engine has been developed. It is desired to determine if the new engine ("B") is superior to the engine currently in use ("A") in terms of MPG.

A wealth of information has been developed over the last few years concerning the performance of engine A. This information is summarized in the reference dist$^{\underline{n}}$ shown below.




Observed MPG for the engine A

What are some of the factors which produce this normal shaped pattern of variability?

- Differences between engines of type A
- Road conditions
- Environmental conditions
- Driver differences
- Speed differences
- Gas differences

Run several trials with Engine B, then compare the results to the reference dist$^{\underline{n}}$ for engine A. Could the results for B have come from dist$^{\underline{n}}$ A?

Perform a statistical test of hypothesis.

- $H_0$: $\mu_B = \mu_A = 20$; $H_A$: $\mu_B \neq \mu_A$.

  or $H_0$: $\mu_B - \mu_A = 0$, $H_A$: $\mu_B - \mu_A \neq 0$
- Let $\alpha = 0.02$, $\frac{\alpha}{2} = .01$ and $Z_{.99} = 2.33$
- If a sample of n=4 for engine B is (18, 24, 28, 22), it can be found that $\bar{B} = 23$
- Calculate Test Statistic:

$$Z = \frac{\bar{B} - \mu_{\bar{B}}}{\sigma_{\bar{B}}} = \frac{\bar{B} - \mu_{\bar{B}}}{\sigma_B / \sqrt{n}}$$ and note that $\mu_B = \mu_A$ and

$\sigma_B = \sigma_A$ under $H_0$, we can find that $Z_{calc}=1.2$

- *Prob* $(\bar{B} \geq 23) = Prob\ (Z \geq 1.2) = 0.1151$, we can't reject $H_0$;

  or since $Z_{calc} = 1.2 \leq 2.33 = Z_{.99}$ we can't reject $H_0$.

For the above engine comparison problem, we can develop a 98% CI instead of performing a statistical test. It can be easily calculated that:

$$\mu_{LO}, \mu_{H1} = \bar{B} \pm Z_{.99}\, \sigma_{\bar{B}} = 23 \pm (2.33)\left(\frac{5}{\sqrt{4}}\right) = 23 \pm 5.825$$

or with 98% confidence, $17.175 \leq \mu_B \leq 28.825$

Since the mean of Engine A ($\mu_A = 20$) lies on the interval, we can't reject the hypothesis that $\mu_A = \mu_B$, or no difference between A and B is evident. Therefore, there is no need to introduce Engine B.

What is wrong with the experiment we have just performed? Statistical analysis was ok, but the design of the comparative test was poor. Additionally, the character of the historical data may be suspect.

The data used to create the reference dist$^{\underline{n}}$ for Engine A was collected under a variety of conditions. These changing conditions introduced a considerable amount of variability and inflate the variance of the reference dist$^{\underline{n}}$.

For our experiment, we compared an experimental result with an external reference dist$^{\underline{n}}$. A better comparison may be made by comparing the two engines side-by-side within the same experiment. Run a few trials for each engine under as similar condition as possible. The reference dist$^{\underline{n}}$ may then be constructed from the experimental data - internal reference dist$^{\underline{n}}$.

**A new experiment**: Select one engine of each type ("A" and "B") at random. Conduct several MPG trials with each engine using one driver (short time period). Driver selects route and engine to be tested at random.

A:14,22,17,21,26,19
B: 18,23,27,20,24

| (Thursday) Engine A | (Friday) Engine B |
|---|---|
| $\bar{A}$= 19.833 | $\bar{B}$= 22.4 |
| $n_A$ =6 | $n_B$ = 5 |
| $\nu_A$ = 5 | $\nu_B$=4 |
| $s_A^2$ = 17.367 | $s_B^2$ = 12.3 |

Let's assume that the population variance, $\sigma^2$, for the two similar engines (A and B) are equal ($\sigma_A^2 = \sigma_B^2 = \sigma^2$). We have two estimates for this variance, $s_A^2$ and $s_B^2$. Let's combine these to give us a better estimate: $s_{pooled}^2 = \dfrac{s_A^2\nu_A + s_B^2\nu_B}{\nu_A + \nu_B}$.

So, $s_{pooled}^2 = \dfrac{17.367(5) + 12.3(4)}{5 + 4} = 15.115$.

Note that if $\nu_A = \nu_B$, this reduces to: $s^2_{pooled} = \frac{s_A^2 + s_B^2}{2}$

$s^2_{pooled}$ or $s^2_p$ is a weighted average of the independent sample variances.

$H_0$: $\mu_A - \mu_B = 0$, $H_a$: $\mu_B - \mu_A \neq 0$ and $\alpha = 0.02$. To test $H_0$, we will use $\bar{B}$ - $\bar{A}$. What does the dist$^{\underline{n}}$ of "$\bar{B}$ - $\bar{A}$" look like under $H_0$?

Note that $\mu_{\bar{B}-\bar{A}} = E(\bar{B}-\bar{A}) = \mu_{\bar{B}} - \mu_{\bar{A}} = 0$ under $H_0$

$$\sigma^2_{\bar{B}-\bar{A}} = Var(\bar{B}-\bar{A}) = Var(\bar{B}) + Var(\bar{A}) = \sigma^2_{\bar{B}} + \sigma^2_{\bar{A}} = \frac{\sigma^2_B}{n_B} + \frac{\sigma^2_A}{n_A}$$

If $\sigma^2 = \sigma^2_A = \sigma^2_B$, then $\sigma^2_{\bar{B}-\bar{A}} = \sigma^2\left(\frac{1}{n_B} + \frac{1}{n_A}\right)$. If we use $s^2_p$ as an estimate of $\sigma^2$, then :

$$\hat{\sigma}^2_{\bar{B}-\bar{A}} = s^2_p\left(\frac{1}{n_B} + \frac{1}{n_A}\right) = 15.115\left(\frac{1}{5} + \frac{1}{6}\right) = 5.542.$$

If A & B are normal, then so is $\bar{B}$-$\bar{A}$. At the very least, $\bar{B}$-$\bar{A}$ is approximately normal regardless of the dist$^{\underline{n}}$ of A and B. $\bar{B}$-$\bar{A}$ is a linear combination of 11 observations/responses,

$$\bar{B} - \bar{A} = \frac{B_1 + B_2 + B_3 + B_4 + B_5}{5} - \frac{A_1 + A_2 + A_3 + A_4 + A_5 + A_6}{6}$$

The quantity

$$t_{calc} = \frac{(\bar{B}-\bar{A}) - (\mu_{\bar{B}} - \mu_{\bar{A}})}{s_{\bar{B}-\bar{A}}} = \frac{(\bar{B}-\bar{A}) - \mu_{\bar{B}-\bar{A}}}{s_{\bar{B}-\bar{A}}} \quad \text{is}$$

distributed according to t distribution with $\nu$=9. In this case we have, $t_{calc} = \frac{2.567 - 0}{2.3542} = 1.09$.

**Statistical Decision:**

Since $t_{calc} < t_{9,0.99}$, it appears as if the observed difference $(\bar{B} - \bar{A})$, is fairly typical. No reason to reject $H_0$: $(\mu_B = \mu_A)$ or $(\mu_B = \mu_A = 0)$ Engine "B" was not demonstrated to be superior to Engine "A". There is no reason to adopt "B".

**Confidence Interval**

For above engine problem, we can also develop a confidence interval for the true mean difference $\mu_B$-$\mu_A$. The 100(1-α)% C.I. for a random variable with variance unknown is :

$$STAT \pm t_{\nu,\, 1-\frac{\alpha}{2}} S_{STAT}$$

where $v$ is the degrees of freedom associated with $S_{STAT}$. In our case, α=0.02, ν=9 and $t_{9,\, 0.99} = 2.821$. The 100(1-α)% C.I. is

$(\bar{B} - \bar{A}) \pm t_{9,\, 0.99} s_{\bar{B} - \bar{A}} = 2.567 \pm 2.821(2.3542)$ or

$-4.07 \le \mu_B - \mu_A \le 9.21$

Since the Confidence Interval include the 0 point, it can't reject the $H_0$ that $\mu_B - \mu_A = 0$.

What's wrong with this experiment? What's good?

- All "A"s run on Thurs., "B"s on Fri. If engines were statistically different, we don't know whether due to a true engine difference or the difference from Thursday to Friday Experiments. It means that Thursday → Friday effect on MPG confounded with Engine's A → B effect. Can these effects be properly interpreted experimentally?
- Day on which each engine's trial is performed should be random. If day selected randomly, then the experiment is valid.
- Fixed a number of factors that varied randomly in the original experiment. So we can reduce the variability of the reference dist$^{\underline{n}}$. This could lead a more sensitive experiment

and comparison.

- If route is selected randomly, the experiment is valid. However, differences between routes will inflate the variability of the reference dist$^{\underline{n}}$. We can't fix the route, because the engine differences may be dependent on the route (Hills - Engine B; Flat - Engine A). We would like to know this. Therefore, we need to somehow filter/block out the route-to-route variation.

**Paired t-test**

All factors but day - weather and route are fixed. Seven engines of each type selected. Seven routes selected - widely different. Each engine type tested on each route - day of test (M or Tu selected randomly)

| Eng. A | Eng. B |
|---|---|
| 14 (Tu) | 17 (M) |
| 21 (M) | 23 (Tu) |
| 17 (Tu) | 21 (M) |
| 22 (Tu) | 26 (M) |
| 15 (M) | 19 (Tu) |
| 20 (Tu) | 23 (M) |
| 24 (M) | 25 (Tu) |
| $\bar{A} = 19$ | $\bar{B} = 22$ |
| $s_A^2 = 14$ | $s_B^2 = 10.333$ |

$$s_P^2 = \frac{\nu_A s_A^2 + \nu_B s_B^2}{\nu_A + \nu_B} = \frac{6(14) + 6(10.333)}{6 + 6} = 12.167$$

The $S_P^2$ can be viewed as an estimate of the common population variance.

$$s_{\bar{B} - \bar{A}}^2 = s_P^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right) = (12.167)\left(\frac{1}{7} + \frac{1}{7}\right) = 3.473$$

and the 98% confidence interval is $(\bar{B} - \bar{A}) \pm t_{12,\,0.99} s_{\bar{B}-\bar{A}} = 3 \pm 2.681(1.8645)$ or,

$-2 \le \mu_B - \mu_A \le 8$

The C.I. includes 0, so we can't reject $H_0 : (\mu_B - \mu_A = 0)$. The reason may be $s_P^2$ and $s_{\bar{B}-\bar{A}}^2$ are inflated by route - to - route differences.

If we compare MPG's for each route, $B_i - A_i$, then this difference "blocks" out the route-to-route effect on the variability. Since within a pair we randomly assigned the day, any effect due to the day (weather) is randomly distributed so as not to bias the experimental results.

| Route | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| A | 14 | 21 | 17 | 22 | 15 | 20 | 24 |
| B | 17 | 23 | 21 | 26 | 19 | 23 | 25 |
| d=B-A | 3 | 2 | 4 | 4 | 4 | 3 | 1 |

$H_0 : (\mu_d = 0)$, $H_a : \mu_d \neq 0$ , $\alpha = 0.02$ , $\frac{\alpha}{2} = 0.01$

$d = 3$, $s_d^2 = 1.333$, $\nu_d = 6$, and $s_{\bar{d}} = \frac{S_d}{\sqrt{n}} = 0.4364$.

Compare $S_d$ with $s_{\bar{B}-\bar{A}}$, we can see how $s_{\bar{B}-\bar{A}}$ was inflated.

**Confidence Interval:**

$\bar{d} \pm t_{6,.99}\, s_{\bar{d}} = 3 \pm 3.143\ (.4364)$ or, $1.628 \le \mu_d \le 4.372$ is the 98% C.I.

Since zero is not within the C.I., we reject $H_0$ and accept $H_a$ $\mu_d \neq 0$. Strong evidence suggests that $\mu_B \neq \mu_A$, and we should accept Engine B.

**Two Important Concepts:**

Blocking: Identify important factors ahead of time, and design experiment so that comparisons may be made between "treatments" (Engine A vs. Engine B) within a block. Representative variation between blocks should be encouraged.

Randomization:
It is Important for 2 reasons:

- Experimental validity. Do not want any known or unknown factors confounded with the treatments (for example, all A's on Monday and all B's on Tuesday). Randomization was achieved in our paired experiment by tossing a coin, Head: Engine A on Monday (B on Tuesday) Tail: Engine A on Tuesday (B on Monday). Tosses obtained were T,H,T,T,H,T,H.
- Randomization DIST$^{\underline{n}}$ idea. Under $H_0$, the label assigned to an engine (A or B) within a route is arbitrary. If we were to switch the labels the sign on the measurement of d would switch. The observed sequence of 7d's is one of $2^7$ possible sequences that could be obtained by switching labels. Each sequence of 7 d's would produce a different $\bar{d}$. The complete set of $\bar{d}$'s is given by: .

$$\bar{d} = \frac{\pm3\pm2\pm4\pm4\pm4\pm3\pm1}{7}$$

The dist$^{\underline{n}}$ of $\bar{d}$'s defined by this equation is referred to as the randomization dist$^{\underline{n}}$.

We observe that our $\bar{d} = 3$ is larger than all other $\bar{d}$'s. We may refer our calculated $\bar{d}$ to the random dist$^{\underline{n}}$. Significance levels for the rand dist$^{\underline{n}}$ closely match those of t dist$^{\underline{n}}$ with appropriate degrees of freedom. Therefore, provided we randomize the experiment, we may use the randomization dist$^{\underline{n}}$ without making any assumptions about the parent population.

**Rules to live by**:

- "Block what you can, randomize to handle factors which are not blocked"
- Identify "important" factors in the experiment.
- Fix as many nuisance variables as possible
- Block factors which can be controlled but not fixed. Or, we don't wish to fix them.
- Randomize the experiment. To handle unknown factors or factors that aren't blocked or fixed.
  - Ensure experimental validity
  - Guarantee approximate validity of t-test.