Compare two datasets

$X_1, \ldots, X_N \overset{d}{=} \text{Normal}(\mu_X, \sigma^2)$ (Treatment group)

$Y_1, \ldots, Y_M \overset{d}{=} \text{Normal}(\mu_Y, \sigma^2)$ (Control group)

independent

Applications - Medicine, (treatment)

Effect of treatment $:= \mu_X - \mu_Y$

Estimate effect: $\bar{X} - \bar{Y}$

Observation: $\bar{X} - \bar{Y} \overset{d}{=} \text{Normal}\left(\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n} + \frac{1}{n}\right)\right)$

- From facts in Probability about i.i.d Normals

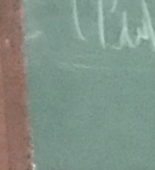
Earlier techniques

σ^2 - known

$100(1-\alpha)\%$ CI for $\mu_X - \mu_Y$

$$\left(\bar{X} - \bar{Y} - z_{\alpha/2}\sigma\sqrt{\frac{1}{n}+\frac{1}{m}} \; ; \; \bar{X} - \bar{Y} + z_{\alpha/2}\sigma\sqrt{\frac{1}{n}+\frac{1}{m}} \right)$$

where $P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$; $Z \overset{d}{=} N(0,1)$



σ^2 - unknown

$$S_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

$$S_Y^2 = \frac{\sum_{i=1}^{m}(Y_i - \bar{Y})^2}{m-1}$$

Pooled Estimate of sample variance

$$S^2_{Pooled} := \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

$$S_{\bar{X}-\bar{Y}} := \sqrt{\frac{1}{n}+\frac{1}{m}}\, S_{Pooled}$$

S^2_{pooled}

= weighted average of the two sample variances
- weights proportional to degrees of freedom of each
- Corrects y m or n is larger

Fact: (Probability)

$$\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{1}{n}+\frac{1}{m}}\,S_{pooled}} = \frac{d}{=} t_{m+n-2}$$

$100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$

$$\left(\bar{X}-\bar{Y} - t^{(\alpha/2)}_{m+n-2}\,S_{\bar{X}-\bar{Y}},\ \bar{X}-\bar{Y} + t^{(\alpha/2)}_{m+n-2}\,S_{\bar{X}-\bar{Y}}\right)$$

$$P\left(T > t^{(\alpha/2)}_{m+n-2}\right) = \frac{\alpha}{2}$$

Testing

Null hypothesis : $\mu_X = \mu_Y$

Fix level of significance α

Alternate hypothesis

	$\mu_X > \mu_Y$	$\mu_X < \mu_Y$	$\mu_X \neq \mu_Y$
p-value Computation	$P\left(T > \frac{(\bar{X}-\bar{Y})}{S_{\bar{X}-\bar{Y}}}\right)$	$P\left(T < \frac{\bar{X}-\bar{Y}}{S_{\bar{X}-\bar{Y}}}\right)$	$P\left(\lvert T\rvert > \frac{\bar{X}-\bar{Y}}{S_{\bar{X}-\bar{Y}}}\right)$

If p-value $< \alpha$ then we will reject null hypothesis.

Analys

Generalis

Model

Y_{ij}

D

Analysis of Variance – (Method)

Generalisation: I – treatments

J – sample size n each treatment group

Model: measure effect of response to treatment

Y_{ij} $\quad i \in \{1, \dots I\}$, $j \in \{1, \dots J\}$

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Differential effect

$\sum_{i=1}^{I} \alpha_i = 0$

$\varepsilon_{ij} \overset{d}{=} N(0, \sigma^2)$

independence

If $\alpha_i =$ treat same

$\alpha_{i_1} - \alpha_{i_2}$

$(I = 2$

$\bar{Y} =$

Test:

If $\alpha_i = 0 \quad \forall i \in \{1 \dots I\}$ then all treatments (are same) have the same expected response effect

$\alpha_{i_1} - \alpha_{i_2}$:= difference in expected response across treatments i_1 & i_2

($I = 2$ → comparing two datasets of same size)

$$\bar{Y} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} Y_{ij}}{I\,J}$$

overall

$$\bar{Y}_{i\cdot} = \frac{\sum_{j=1}^{J} Y_{ij}}{J}$$

within treatment

Test: Null Hypothesis $\alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$