Handout: **Two Sample Hypothesis Testing And Inference For Difference In Means**

Hypothesis Testing

I. *Two independent samples from Normal distributions.*

Suppose $X_1, ..., X_{n_1}$ is an independent sample from $Normal(\mu_1, \sigma_1^2)$ distribution. Independently of the first sample, suppose $Y_1, ..., Y_{n_2}$ is an independent sample from $Normal(\mu_2, \sigma_2^2)$ distribution (possibly different from the first one):

	Group1	Group2	Known?
Mean	μ_1	μ_2	Unknown
Variance	σ_1^2	σ_2^2	Either
Data	$X_1, ..., X_{n_1}$	$Y_1, ..., Y_{n_2}$	Known
Sample size	n_1	n_2	Known
Sample Mean	$m_1 = \bar{X}$	$m_2 = \bar{Y}$	Known
Standard Deviation	SD_1	SD_2	Known
Distribution	$Normal(\mu_1, \sigma_1^2)$	$Normal(\mu_2, \sigma_2^2)$	Assumed

Reasonable estimate for the difference of the population means $\mu_1 - \mu_2$ is

$$m_1 - m_2 = \bar{X} - \bar{Y}.$$

Note that

$$E(m_1 - m_2) = \mu_1 - \mu_2$$

and

$$SE(m_1 - m_2) = \sqrt{Var(m_1 - m_2)} = \sqrt{Var(m_1) + Var(m_2)} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}.$$

for independent samples $X_1, ..., X_{n_1}$ and $Y_1, ..., Y_{n_2}$.

For testing

$$H_0: \quad \mu_1 = \mu_2 \tag{1}$$
$$\text{against}$$
$$H_1: 1)\mu_1 \neq \mu_2 \text{ or}$$
$$2)\mu_1 < \mu_2 \text{ or}$$
$$3)\mu_1 > \mu_2$$

use

$$\text{test statistics } d^{\star}_{obt} = \frac{m_1 - m_2}{SE(m_1 - m_2)}$$

which follows some distribution d^*.

Theorem. Under the above assumptions about the two samples $X_1, ..., X_{n_1}$ and $Y_1, ..., Y_{n_2}$, for testing test H_0: $\mu_1 - \mu_2 = 0$ at α - significance level

vs

1) H_1: $\mu_1 - \mu_2 \neq 0$. Reject H_0 if $|d^*_{obt}| \geq d^*_{crit}(\alpha/2)$

2) H_1: $\mu_1 - \mu_2 < 0$. Reject H_0 if $d^*_{obt} \leq -d^*_{crit}(\alpha)$

3) H_1: $\mu_1 - \mu_2 > 0$. Reject H_0 if $d^*_{obt} \geq d^*_{crit}(\alpha)$

Computation of $SE(m_1 - m_2)$ and choice of distribution d^*:

1. σ_1 and σ_2 are known
   $SE(m_1 - m_2) = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$
   Test statistics $d^*_{obt} = z_{obt} = \frac{m_1 - m_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$ follows standard Normal distribution z.

2. σ_1 and σ_2 are unknown, but n_1 and n_2 are large (≥ 30)
   In this case can omit Normality assumption.
   $SE(m_1 - m_2) = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} \approx \sqrt{SD_1^2/n_1 + SD_2^2/n_2}$ and
   test statistics $d^*_{obt} = z_{obt} = \frac{m_1 - m_2}{\sqrt{SD_1^2/n_1 + SD_2^2/n_2}}$ approximately follows standard Normal distribution z.

3. σ_1 and σ_2 are unknown, and n_1 and n_2 are not large enough (≤ 30)

   a) $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (unknown)
   $SE(m_1 - m_2) = \sqrt{\sigma^2(1/n_1 + 1/n_2)}$ with
   pooled estimate of σ^2: $\sigma^2_{pool} = \frac{(n_1-1)SD_1^2+(n_2-1)SD_2^2}{n_1+n_2-2}$ and
   test Statistics $d^*_{obt} = t_{obt} = \frac{m_1-m_2}{\sqrt{\sigma^2_{pooled}(1/n_1+1/n_2)}}$ follows t distribution with $df = n_1 + n_2 - 2$ degrees of freedom.

   b) $\sigma_1^2 \neq \sigma_2^2$ are unknown
   $SE(m_1 - m_2) = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} \approx \sqrt{SD_1^2/n_1 + SD_2^2/n_2}$ and
   test Statistics $d^*_{obt} = t_{obt} = \frac{m_1-m_2}{\sqrt{SD_1^2/n_1+SD_2^2/n_2}}$ follows t distribution with degrees of freedom:
   $df = \frac{(SD_1^2/n_1+SD_2^2/n_2)^2}{\frac{(SD_1^2/n_1)^2}{n_1-1}+\frac{(SD_2^2/n_2)^2}{n_2-1}}$ (estimated by MATLAB)
   or alternatively $df \approx \min(n_1 - 1, n_2 - 1)$.

If instead of testing (1), want to test

$$H_0 : \mu_1 - \mu_2 = d \quad \text{against} \quad H_1 : \mu_1 - \mu_2 \neq d (< \text{ or } >) \tag{2}$$

use

$$\text{test statistics } d^*_{obt} = \frac{m_1 - m_2 - d}{SE(m_1 - m_2)}$$

II. *Proportions*

For a random variable X drawn from a $Binomial(n_1, p_1)$ distribution, and an independent random variable Y drawn from a $Binomial(n_2, p_2)$ distribution, let $\bar{p_1} = X/n_1$ and $\bar{p_2} = Y/n_2$. For testing

$$\begin{aligned} H_0 :\ & p_1 = p_2 \\ & \text{against} \\ H_1 :\ & 1) p_1 \neq p_2 \text{ or} \\ & 2) p_1 < p_2 \text{ or} \\ & 3) p_1 > p_2 \end{aligned} \tag{3}$$

use test statistics $d^{\star}_{obt} = \frac{\bar{p_1}-\bar{p_2}}{SE(\bar{p_1}-\bar{p_2})}$.

Since for a $Binomial(n,p)$ random variable X and $\bar{p} = X/n$, $Var(\bar{p}) = \frac{p(1-p)}{n}$,

$$SE(\bar{p_1}-\bar{p_2}) = \sqrt{Var(\bar{p_1}-\bar{p_2})} = \sqrt{Var(\bar{p_1})+Var(\bar{p_2})} = \sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

Then

$$\text{test statistics } d^{*}_{obt} = z_{obt} = \frac{\bar{p_1}-\bar{p_2}}{\sqrt{\frac{\bar{p_1}(1-\bar{p_1})}{n_1}+\frac{\bar{p_2}(1-\bar{p_2})}{n_2}}}$$

has approximately Normal z distribution if n_1p_1, $n_1(1-p_1)$, n_2p_2, and $n_2(1-p_2)$ ≥ 10.

III. *Dependent samples (Paired data)*

For paired measurements $(X_1,Y_1),...,(X_n,Y_n)$ (eg., measurements "before" and "after") previous theory does not hold.

Sample $X_1,...,X_n$ (an independent sample from $Normal(\mu_1,\sigma_1^2)$ distribution)is not independent of sample $Y_1,...,Y_n$ (an independent sample from $Normal(\mu_2,\sigma_2^2)$ distribution), then

$$SE(m_1-m_2) = \sqrt{Var(m_1-m_2)} = \sqrt{Var(m_1)+Var(m_2)-Cov(m_1,m_2)} \neq \sqrt{Var(m_1)+Var(m_2)}$$

since $Cov(m_1,m_2) \neq 0$ for non-independent data!

In such case, testing (1) is equivalent to testing one-sample hypothesis for data $D1(=X_1-Y_1)$, ... , $D_n(=X_n-Y_n)$:

$$H_0: \quad \mu_D = 0 \tag{4}$$

against

H_1 :1)$\mu_D \neq 0$ or

2)$\mu_D < 0$ or

3)$\mu_D > 0$.

Confidence Intervals

For a test statistics $d^*_{obt} = \frac{m_1 - m_2}{SE(m_1 - m_2)}$ we reject H_0 if $|d^*_{obt}| \geq d^*_{crit}(\alpha/2)$. If

$$-d^*_{crit}(\alpha/2) < d^*_{obt} < d^*_{crit}(\alpha/2)$$

we conclude that evidence against H_0 is not statistically significant at α - significance level.

Conficence interval for $\mu_1 - \mu_2$ is computed by inverting non-rejection region

$$-d^*_{crit}(\alpha/2) < \frac{m_1 - m_2}{SE(m_1 - m_2)} < d^*_{crit}(\alpha/2)$$
$$-d^*_{crit}(\alpha/2)SE(m_1 - m_2) < m_1 - m_2 < d^*_{crit}(\alpha/2)SE(m_1 - m_2)$$

with $(1 - \alpha)100\%$ confidence interval for $\mu_1 - \mu_2$:

$$((m_1 - m_2) - d^*_{crit}(\alpha/2)SE(m_1 - m_2);\ (m_1 - m_2) + d^*_{crit}(\alpha/2)SE(m_1 - m_2))$$