Inference Testing II

- The difference of two means ($H_0 : \mu_1 = \mu_2$)
  - With large samples

$$T = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

$\cancel{N}$

T

  - Rejection region

$$H_1: \mu_1 > \mu_2, \qquad T \geq Z_\alpha$$
$$H_1: \mu_1 < \mu_2, \qquad T \leq -Z_\alpha$$
$$H_1: \mu_1 \neq \mu_2, \qquad |T| \geq Z_{\frac{\alpha}{2}}$$

한양대학교 HANYANG UNIVERSITY

$H_0 : \mu_1 - \mu_2 = 0$

$H_0 : P_1 - P_2 = 0$

$H_0 : \sigma_1^2 / \sigma_2^2 = 1$

$H_1 : \mu_1 < \mu_2 \Leftrightarrow \mu_1 - \mu_2 < 0$

$\mu_1 > \mu_2 \qquad > 0$

$\mu_1 \neq \mu_2 \qquad \neq 0$

Inference Testing II

- The difference of two means
  - With small samples

$$T = \frac{\overline{x_1} - \overline{x_2}}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}, \; s_p^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 - n_2 - 2}$$

  - Rejection region

$$H_1: \mu_1 > \mu_2, \qquad T \geq t_{(\alpha, n_1+n_2-2)}$$
$$H_1: \mu_1 < \mu_2, \qquad T \leq -t_{(\alpha, n_1+n_2-2)}$$
$$H_1: \mu_1 \neq \mu_2, \qquad |T| \geq t_{(\frac{\alpha}{2}, n_1+n_2-2)}$$

한양대학교 HANYANG UNIVERSITY

Inference Testing II

- The difference of two means
  - Matched cases

$$T = \frac{\bar{D}}{s_D/\sqrt{n}}$$

$$\bar{D} = \frac{\sum_{i=1}^{n}(D_i)}{n} = \frac{\sum_{i=1}^{n}(x_{i1}-x_{i2})}{n}$$

$$s_D^2 = \frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1}$$

  - Rejection region

$$H_1: \mu_1 > \mu_2, \quad T \geq t_{(\alpha, n-1)}$$
$$H_1: \mu_1 < \mu_2, \quad T \leq -t_{(\alpha, n-1)}$$
$$H_1: \mu_1 \neq \mu_2, \quad |T| \geq t_{(\frac{\alpha}{2}, n-1)}$$

한양대학교 HANYANG UNIVERSITY

p	pre - post = d_i	
1	$X_1 - Y_1$	
2	$X_2 - Y_2$	
⋮	⋮	
20	$X_{20} - Y_{20}$	

$$\bar{D} = \frac{\sum_{i=1}^{n} (X_i - Y_i)}{n}$$

$$S^2 = \frac{\sum_{i=1}^{n} (d_i - \bar{D})^2}{n-1}$$

Inference Testing II

```
In [1]: import numpy as np
        import seaborn as sns
        from scipy.stats import ttest_ind, ttest_rel
```

ttest_ind
-. *T-test of two independent samples* 남 vs 여

ttest_rel
-. *T-test of two related samples*

한양대학교 HANYANG UNIVERSITY

Inference Testing II

.ttest_ind(*x, y, equal_var, alternative*)
-. *x, y: two arrays*
-. *equal_var: True or False*
-. *alternative: 'two-sided' or 'less' or 'greater'*

Inference Testing II

- Practice
  - Test if two means are equal or not at $\alpha = 0.05$

```
In [3]: A_group = can[0:10]
        B_group = can[-10:]

        print("A group:",A_group)
        print("B group:",B_group)

A group: [101.8 101.5 102.6 101.  101.8  96.8 102.4 100.   98.8  98.1]
B group: [101.2  99.9  99.1 100.7 100.8 100.8 101.4 100.3  98.4  97.2]
```

Inference Testing II

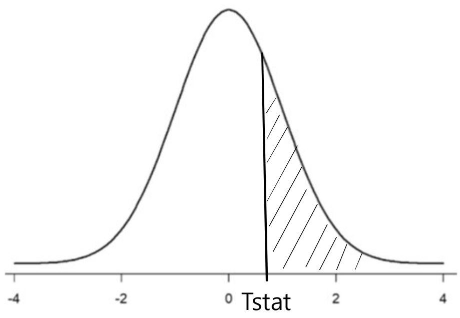
- Solution
  - Test if two means are equal or not at $\alpha = 0.05$

```
In [4]: tstat, pval = ttest_ind(A_group, B_group, equal_var=True, alternative='two-sided')
        print("T stat is ", tstat)
        print("p-value is ",pval)

T stat is  0.6596226981846296
p-value is  0.5178474668321495
```


Inference Testing II

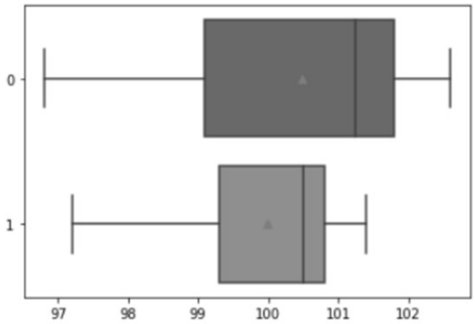
- Result -interpretation in *stats.*
  - *alternative* = ‘two-sided’
    -. Right side area x 2 then ??
  - *alternative* = ‘less’
    -. Left side area of *Tstat*
  - *alternative* = ‘greater’
    -. Right side area of *Tstat*



Inference Testing II

- Solution (continued)
  - Reject ?

```
In [20]: sns.boxplot(data=[A_group, B_group], orient='h', showmeans=True)
Out[20]: <AxesSubplot:>
```



한양대학교 HANYANG UNIVERSITY

Inference Testing II

.ttest_rel(*x, y, alternative*)
-. *x, y : two arrays*
-. *alternative: 'two-sided' or 'less' or 'greater'*
ex) 'less': x is less than y

한양대학교 HANYANG UNIVERSITY

Inference Testing II

- Practice
  - There are midterm and final exam scores of15 students. Test if the final score is higher than midterm score at $\alpha = 0.05$

```
In [10]: midterm = np.array([80,73,70,60,88,84,65,37,91,98,52,78,40,79,59])
         final = np.array([82,71,95,69,100,71,75,60,95,99,65,83,60,86,62])
```

Inference Testing II

- Solution

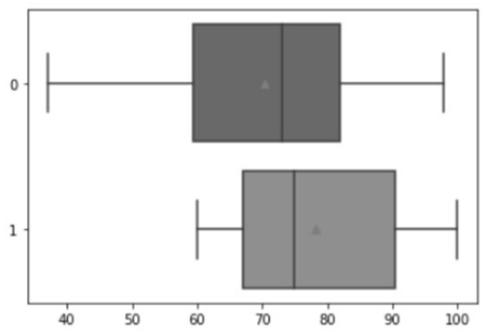
```
In [21]: tstat, pval = ttest_rel(midterm, final, alternative = 'less')
         print("T stat is ", tstat)
         print("p-value is ",pval)

T stat is  -3.093705670004429
p-value is  0.003965461614513267
```

Inference Testing II

- Solution
  - Then??

```
In [23]: sns.boxplot(data = [midterm, final], orient='h', showmeans=True)
Out[23]: <AxesSubplot:>
```



Inference Testing II

- The difference of two proportions ($H_0 : p_1 = p_2$)

$$T = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\hat{p}(1-\hat{p})} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1) \qquad \hat{p} = \frac{(x_1 + x_2)}{(n_1 + n_2)}$$

  - Rejection region

$$H_1: p_1 > p_2, \qquad T \geq Z_\alpha$$
$$H_1: p_1 < p_2, \qquad T \leq -Z_\alpha$$
$$H_1: p_1 \neq p_2, \qquad |T| \geq Z_{\frac{\alpha}{2}}$$

$\frac{X}{n_1}$ $\frac{Y}{n_2}$

$H_0: p_1 - p_2 \Leftrightarrow \boxed{p_1 - p_2 = 0}$

$H_1: \; < \; \Leftrightarrow \; <0$

$> \quad >0$

$\neq \quad \neq 0$

$$\hat{p} = \frac{X+Y}{n_1+n_2}$$

Inference Testing II

- Practice
  - In A class, 4 students are absent among 25 students, and 6 students were absent among 20 students in B class. Test if the absence proportion of A class is less than B class at $\alpha = 0.05$

Inference Testing II

- Solution

In [32]:

```
from scipy.stats import norm

def two_prop(x, n1, y, n2, alternative):
    phat1 = x/n1
    phat2 = y/n2
    phat = (x+y)/(n1+n2)

    tstat = (phat1-phat2)/(np.sqrt(phat*(1-phat))*np.sqrt(1/n1 + 1/n2))
    if alternative == 'less':
        pval = norm.cdf(tstat)
    elif alternative =='greater':
        pval = 1- norm.cdf(tstat)
    else:
        pval = 2*(1- norm.cdf(tstat))
    return tstat, pval
```

Inference Testing II

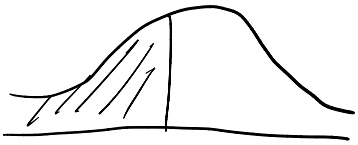
- Solution

```
In [33]: tstat, pval = two_prop(4,25, 6,20, 'less')

         print("T stat is, ", tstat)
         print("P-value is ", pval)

T stat is,  -1.1224972160321824
P-value is  0.13082554529411833
```

$0.05 < 0.13$



0.13

Inference Testing II

- The equality of two variances ($H_0 : \sigma_1^2 = \sigma_2^2$)

$$T = \frac{s_1^2}{s_2^2} \sim F_{(n_1-1, n_2-1)}$$

- Rejection region

$$H_1: \sigma_1^2 > \sigma_2^2, T \geq F_{(\alpha, n_1-1, n_2-1)}$$

$$H_1: \sigma_1^2 < \sigma_2^2, T \leq F_{(1-\alpha, n_1-1, n_2-1)} = \frac{1}{F_{(\alpha, n_2-1, n_1-1)}}$$

$$H_1: \sigma_1^2 \neq \sigma_2^2, T \leq F_{\left(\frac{\alpha}{2}, n_2-1, n_1-1\right)} \; or \; T \geq F_{\left(\frac{\alpha}{2}, n_1-1, n_2-1\right)}$$

Inference Testing II

- Practice
  - Test if the two variances are equal at $\alpha = 0.05$

```
In [3]: A_group = can[0:10]
        B_group = can[-10:]

        print("A group:",A_group)
        print("B group:",B_group)

        A group: [101.8 101.5 102.6 101.   101.8  96.8 102.4 100.   98.8  98.1]
        B group: [101.2  99.9  99.1 100.7 100.8 100.8 101.4 100.3  98.4  97.2]
```


Inference Testing II

- Solution
  - F-distribution can be called from *stats*

```
In [27]: from scipy.stats import f

         def test_var2(x, y, alternative):
             tstat = np.var(x, ddof=1)/np.var(y, ddof=1)
             df1 = len(x)-1
             df2 = len(y)-1
             if alternative == 'less':
                 pval = f.cdf(tstat,df1,df2)
             elif alternative == 'greater':
                 pval = 1-f.cdf(tstat,df1,df2)
             else :
                 pval = 2*(1-f.cdf(tstat,df1,df2))
             return tstat, pval
```

Inference Testing II

- Solution
  - we can say that two variances are equal at $\alpha = 0.05$

```
In [28]: tstat, pval = test_var2(A_group, B_group, alternative='two-sided')
         print("T stat is, ", tstat)
         print("P-value is ", pval)

T stat is,  2.138625880067981
P-value is  0.27285489700952237
```


Inference Testing II

stats.bartlett(*x*, *y*) 시험X.
-. *x, y : two arrays*

-. *Return Bartlett test result, which is known to be less sensitive to normality*

Inference Testing II

- Using *Barlett's* test
  - Still the same!

In [30]:
```
from scipy import stats

stats.bartlett(A_group, B_group)
```

Out [30]:
```
BartlettResult(statistic=1.203168038877362, pvalue=0.272689407591583)
```

한양대학교 HANYANG UNIVERSITY

Inference Testing II

- Sample size
  - Limit of error (d)
    : Half-length of the confidence interval
  - Sample size
    : the confidence interval of the mean can be used…!

$$n = \left(Z_{\frac{\alpha}{2}} \cdot \frac{s}{d}\right)^2$$

ppf.

한양대학교 HANYANG UNIVERSITY