## 6.4 Test for population means/proportions

### 6.4.1 Test for one population

#### Population mean

- $H_0 : \mu \leq \mu_0, (\mu = \mu_0, \mu \geq \mu_0)$, vs $H_a : \mu > \mu_0, (\mu \neq \mu_0, \mu < \mu_0)$;
- Test statistic: $T = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$;
- Assume the normality of the data, then under the null $T \sim T_{n-1}$.
- Rejection region: $R = \{T > t_\alpha\}, (R = \{|T| > t_{\alpha/2}\}, R = \{T < -t_\alpha\})$.
- P-value: $P(T_{n-1} > T), (2 * P(T_{n-1} > |T|), P(T_{n-1} < T))$.

**Example 6.4.1.** *Suppose you are the manager of Speedy Oil Change which claims that it will change the oil in customers' cars in less than 30 minutes. Further suppose that several complaints have been filed from customers stating that their oil change took longer than 30 minutes and upper-level management at Speedy Oil Change headquarters has requested that you investigate the complaints. To begin your investigation, you monitor 36 oil changes performed Speedy Oil Change and record the time each customer waited for the oil change to be completed. The number of minutes to complete each of the 36 oil changes is reported below.*

#### Population proportion

**Example 6.4.2.** *Drug X is administered to 100 patients with a particular disease, 55 improve. Test whether this drug is better than drug Y, which is known to produce improvement in 50% of patients.*

- $H_0 : p \leq p_0, (p = p_0, p \geq p_0)$, vs $H_a : p > p_0, (p \neq p_0, p < p_0)$;
- Test statistic: $Z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/\sqrt{n}}}$;
- Under the null $Z \sim N(0,1)$.
- Rejection region: $R = \{Z > z_\alpha\}, (R = \{|Z| > z_{\alpha/2}\}, R = \{Z < -z_\alpha\})$.
- P-value: $P(N(0,1) > Z), (2 * P(N(0,1) > |T|), P(N(0,1) < T))$.

**Sample size calculation**

For hypothesis testing, we are interested in controlling the specified Type II error probability $\beta$ when the unknown parameter being tested is a distance $\delta$ from the null hypothesized value. For a one sample mean test with one-sided testing, the sample size is

$$n = \sigma^2(z_\alpha + z_\beta)^2/\delta^2;$$

for the two-sided test, the sample size is

$$n = \sigma^2(z_{\alpha/2} + z_\beta)^2/\delta^2;$$

For the test of the proportion, the sample size for one-sided testing is

$$n = \frac{p_1(1-p_1)(z_\alpha + z_\beta)^2}{(p_1 - p_0)^2}.$$

**Example 6.4.3.** *We are interested in determining whether the mean volume of fluid delivered to patients during a particular type of neurosurgery is at least 50ml greater than 1500ml. If so then the surgeons need to think about modifying the procedures to reduce the large fluid volumes. Our null hypothesis is that the mean volume, $\mu$, is 1500 ml. We cannot fule out $\mu < 1500$. We believe $\sigma^2$ is roughly 10,000 ml$^2$. Set $\alpha = 0.05$.*

(a) What power will we have to detect a difference in the mean of 50 ml if we use a sample size of 10 patients?

(b) How many patients will we need to detect a difference of 50 ml with at least 80% power?

### 6.4.2 Test for two populations

#### Two independent samples

When there exists two populations, one wants to test for the means of these two populations. The hypothesis can be written as

$$H_0 : \mu_1 - \mu_2 = \delta_0, H_a : \mu_1 - \mu_2 \neq \delta_0.$$

One can also test for the left-sided and right-sided hypothesis. There are two situations to consider with independent samples. When the populations may be assumed to have a common unknown variance $\sigma$, the test statistic is

$$t = \frac{\bar{y}_1 - \bar{y}_2 - \delta_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

$$s_p = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Under the null hypothesis, $t \sim T_{n_1+n_2-2}$.

When the variance are not the same, the test statistic is

$$t = \frac{\bar{y}_1 - \bar{y}_2 - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Under the null hypothesis, $t \sim T_{sat}$ where the degrees of freedom is determined by the Satterthwaite approximation (Welch option),

$$sat = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1-1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2-1}}.$$

Which one should we use? One can test the variance component first.

We assume here that independent random samples are available from both populations. Let $\sigma_1^2$ and $\sigma_2^2$ are the variance of two populations. $s_1^2$ and $s_2^2$ are the sample variances. Consider the hypothesis

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1, H_a : \frac{\sigma_1^2}{\sigma_2^2} \neq 1.$$

The test statistic is

$$F = \frac{s_1^2}{s_2^2}.$$

Under the null hypothesis,

$$F \sim F_{n_1-1,n_2-1}.$$

**Example 6.4.4** (High-throughput data: microarray dataset [Efron(2010)]). *Classic statistical methods were fashioned for problems where the number of parameters is much smaller than the sample size. Current high-throughput scientific technology tends to produce just the opposite situation; modern equipment may permit thousands of measurements on a single individual.*

*Microarrays offer the iconic example. Here $x_j$ is a vector of genetic expression measurements on subject $j$, one for each of $p$ genes, where $p$ is typically several thousand. In the prostate cancer data ([**?**]), there are $p = 6033$ genes measured on each of n=102 men, $n_1 = 50$ healthy controls and $n_2 = 52$ prostate cancer patients. One task is to identify which genes are differentially expressed between the control and treatment.*

(a). Is $\sigma_{1g} = \sigma_{2g}$?

(b). Test if $\mu_{1g} = \mu_{2g}$.

(c). What if we have many genes? Which ones are important?

### Matched pairs

**Example 6.4.5** (English vs Greek)**.** *One wants to investigate whether native English speakers find it easier to learn Greek than native Greek speakers learning English. Thirty-two sentences are written in both languages. Each sentence is scored according to the quantity of errors made by an English speaker learning Greek and by a Greek speaker learning English. It is desired to compare the mean scores of the two groups.*

Let $y_{1j}$ be the sample of $n$ items from the first population, having mean $\mu_1$ and $y_{2j}$ be the sample of $n$ items from the second population, having mean $\mu_2$. Let $d_j = y_{1j} - y_{2j}$ and $\bar{d}$ and $s_d^2$ be the sample mean and variance of the differences.

$$H_0 : \mu_1 - \mu_2 = \delta_0, H_a : \mu_1 - \mu_2 \neq \delta_0.$$

Test static:

$$T = \frac{\bar{d} - \delta_0}{s_d/\sqrt{n}}.$$

Under the null hypothesis, $T \sim T_{n-1}$.

## 6.5 $\chi^2$ test

### 6.5.1 $\chi^2$ goodness-of-fit test

The $\chi^2$ distribution may be used to conduct goodness-of-fit test, i.e.

$$H_0 : \text{ the data are from a specified population },$$

vs

$$H_1 : \text{ the data are from some other distribution.}$$

This test is very useful for the categorical data.

**Example 6.5.1** (Market Share). *Company A has recently conducted aggressive advertising campaigns to maintain and possibly in- crease its share of the market for fabric softener. Their main competitor, Company B, has 40% of the market and a number of other competitors acount for the remaining 15%. To determine whether the market shares changed after the advertising campaign, the marketing manager for Company A solicited the preferences of a random sample of 200 customers of fabric softener. Of the 200 customers, 102 indicated a preference for Company A's product, 82 prefereed Company B's and the remaining 16 preerred the products of one of the competitors. Can the analyst infer at the 5% significance level that customer preferences have changed from their levels before the advertising campaigns were launched?*

**Goal:** Test if the customer preferences have changed from their levels before the campaigns.

**Set the hypothesis:**

1. $H_0 : p_1 = p_{10}, p_2 = p_{20}, \cdots, p_k = p_{k0}$;

2. $H_1$ : At least one $p_i$ is not equal to its specified value.

**Test Statistic:**
Chi-squared goodness-of-fit test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i},$$

where $e_i = np_{i0}$.

**(Asymptotic) sampling distribution**
When the sample size is large enough, then under the null hypothesis, the test statistic follows a $\chi^2_{k-1}$.

**Solution:**

**Example 6.5.2** (Teenage suicide example)**.**

### 6.5.2 Contingency table

**Example 6.5.3. Exit polls in Ohio**
*After the polls close on Election Day, networks compete to be the first to predict which candidate will win. The predictions are based on counts in certain precincts and on exit polls. Exit polls are conducted by asking random samples of voters who have just exited the polling booth for which candidate they voted. In addition to asking for whom they voted, respondents are asked a variety of other questions about gender, age, education, and etc.*

Let $X$ be the candidate that the voters vote, $Y$ be the gender of the voters, and $Z$ be the education level.

1. Are $X$ and $Y$ independent?

2. Are $X$ and $Z$ independent?

**Goal:** Test if the two random variables are independent.

**Set the hypothesis:**

1. $H_0$ : the two variables are independent;

2. $H_1$ : the two variables are dependent.

**Test Statistic:**
Chi-squared goodness-of-fit test statistic is

$$\chi^2 = \sum_{i,j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}},$$

where $e_{ij} = \frac{\text{Row i total * Row j total}}{\text{sample size}}$.
**(Asymptotic) sampling distribution**
When the sample size is large enough, then under the null hypothesis, the test statistic follows a $\chi^2$ with degrees of freedom $(\#Row - 1)(\#Col - 1)$.