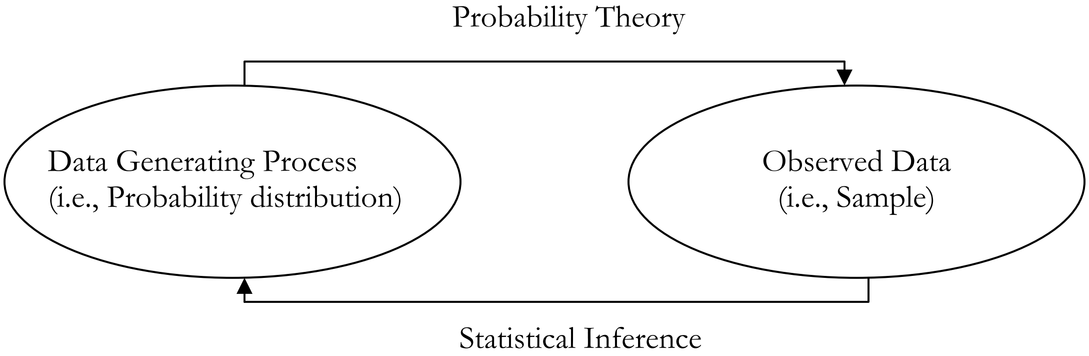
# Handout #4

Title: FAE                                                                 Spring/2015
Course: Econ 368/01                                        Instructor: Dr. I-Ming Chiu

This handout summarizes chapter 3~4 from the reference PE. Relevant reading (detailed ones) can be found in chapter 6, 13, 14, 19, 23, and 25 from MPS. Chapter 6 and 7 from BR also provide information about what covered in this handout.

**Statistical Inference**



Population vs. Sample

Who is to blame for government shutdown? According to an ABC News/Washington Post poll, 63% of Americans disapprove the way GOP handling the budget impasse[1]. What is the average annual household income in the U.S. in 2014? What is the current 30 year fixed mortgage rate? What's the crime rate in Camden County in 2014? We're living in a world with all kinds of different figures and they provide information to help us make better decisions. How do we get these figures? Are they accurate? How do we turn data into useful information? This handout focuses on statistical inferences.

Given that the size of the population we want to study is often huge, it's impractical to study the population features (i.e. mean, variance, proportion, quantiles, etc) by conducting censuses. Instead, a small portion of the population is selected for studying the characteristics of the population. For example, the Census Bureau on behalf of Bureau of Labor conduct survey on 60,000 households monthly in order to find the job market conditions and estimate the unemployment rate. However, this method may introduce uncertainty to our analysis. This uncertainty is termed "sampling variation". For example, if another 60,000 households are chosen and surveyed, the estimate of unemployment rate can be different. How we handle this uncertainty to make sure the information obtained is trustworthy, in terms of probabilistic statements, will be explained in this handout.

[1] http://www.usatoday.com/story/news/politics/2013/09/30/government-shutdown-blame-republicans-polls/2897197/

**Some important usage of Normal, $\chi^2$, t and F distribution in statistical inference**

Suppose a sample of $X_i$ (i=1…n) are drawn from a $N(\mu, \sigma^2)$ RV, then without proof:

$$Z_i = \frac{X_i - \mu}{\sigma} \sim N(0, 1) \text{ for all i (1)}$$

$$\frac{s^2(n-1)}{\sigma^2} \sim \chi^2_{n-1} \qquad (2)$$

$$\frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1} \qquad (3)$$

$$\frac{\chi^2_m / m}{\chi^2_n / n} \sim F_{m,n} \qquad (4)$$

Where

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

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

<u>Theorem</u>: Suppose a random sample of size n is drawn from a population with mean $\mu$ and variance $\sigma^2$, then the sample average $\overline{X}$ will have the following property:

$E(\overline{X}) = \mu$ & $V(\overline{X}) = \frac{\sigma^2}{n}$ , Where $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ (Can you prove both properties?)

Suppose we also know the population has a normal distribution, then $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.

<u>Theorem: (Weak) Law of Large Numbers</u>

Let $X_1, X_2, \ldots, X_n$ be an iid sequence of random variables and $E(X_i) = \mu$, let $S_n = \sum_{i=1}^{n} X_i$ .

$\Rightarrow \frac{S_n}{n} \to \mu$, as $n \to \infty$

Alternatively, the above statement can be written as

$\Rightarrow P(\left|\frac{S_n}{n} - \mu\right| < \varepsilon) \to 1$, as $n \to \infty$, $\forall\ \varepsilon > 0$. (e.g. Binomial RVs using R)

Simulation: Often we don’t know the features about a population and it makes the inference job challenging. However, we can use simulated data to study the linkage between population features and the corresponding sample characteristics.

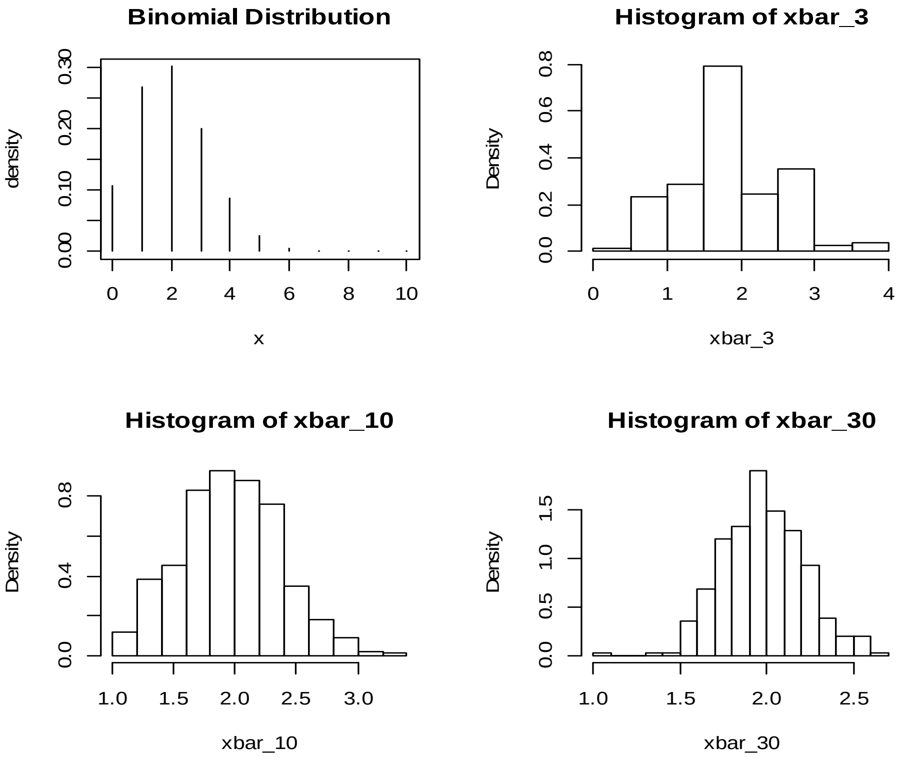
e.g. generate a random sample from a known random variable and study its properties.

The Central Limit Theorem

If large samples (in practice, the size n of each sample is *no less than 30*) are randomly selected from a population with mean $\mu$ and variance $\sigma^2$ , then the sample average $\overline{X}$ will have the following property regardless of the shape of the distribution of the parent population.

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

**Illustrate Central Limit Theorem using Simulation**



Explanation: X is a random variable with Binomial distribution (i.e., X~binom(10, 0.2)). The upper left panel shows what its (theoretical) probability distribution looks like; it is asymmetric and skewed right. We choose repeated samples (i.e., 500 samples) with sample size 3, 10 and 30, respectively. The rest of three histograms show the distribution of the sample average (when n =3, 10 and 30), what do you observe?

The R code used to present CLT can be found in the last part of 368_HD04.txt file. You can resave this part as an R file (e.g. the file extension is .R; e.g., CLT_Binomial.R) and execute it in R as follows: Start R and choose “Source R code” from the pull down menu under “File” in the console, the same diagram as shown above will appear on the screen. You can choose “Open script” to read or modify the code if it’s needed. The details of the code will be gradually covered as we proceed.

**Estimation of Population Characteristics**

**A. Point estimate (population mean, variance, proportion, and covariance)**

Sample Mean: $\overline{X} = \dfrac{\sum_{i=1}^{n} X_i}{n} \Rightarrow E(\overline{X}) = \mu_X$

Sample Variance: $s_X^2 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \Rightarrow E(s_X^2) = \sigma_X^2$

Sample Proportion: $\hat{p} = \dfrac{\sum_{i=1}^{n} X_i}{n}$ (where $X_i = 0$ or $1$) $\Rightarrow E(\hat{p}) = \pi$

Sample Proportion Variance: $\sigma^2_{\hat{p}} = \dfrac{\hat{p}(1-\hat{p})}{n}$

Sample Covariance: $s^2_{X,Y} = \dfrac{\sum_{i=1}^{n}\sum_{j=1}^{m}(X_i - \overline{X}) * (Y_j - \overline{Y})}{n-1}$

Sample Correlation: $r = \dfrac{\text{Sample} \cdot \text{Co var iance}(X, Y)}{s_X s_Y}$

All of these formulas are termed “estimators”.

Notice: (A) There are three methods for obtaining estimators; (i) Least Squares, (ii) Maximum Likelihood and (iii) Method of Moment.
(B) If the expectation of the estimator equals the true parameter, we say the estimator is unbiased. For example, $E(\overline{X}) = \mu_X$.
(C) The other two nice properties for an estimator in addition to unbiasedness are minimum variance and efficiency (compare the properties mentioned in point (B) and (C) to the theorems introduced in page 2 and 3)

## B. Confidence Intervals (CI)

Standardization of $\overline{X} \Rightarrow Z = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$

In practice $\sigma$ is usually unknown, therefore, the sample standard deviation s is used to replace $\sigma$. However, from equation (3) on page 2 we learn that the distribution is no longer standard normal, we should write

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

Use the t distribution[2] table or from R we can find that

$\Pr(-H < t_{n-1} < H$[3]$) = 0.95$

$$\Pr\left(-H < \frac{\overline{X}-\mu}{s/\sqrt{n}} < H\right) = 0.95 \Rightarrow$$

$$\Pr\left(-H*\frac{s}{\sqrt{n}} < \overline{X} - \mu < H*\frac{s}{\sqrt{n}}\right) = 0.95\Rightarrow$$

$$\Pr\left(H*\frac{s}{\sqrt{n}} > \mu - \overline{X} > -H*\frac{s}{\sqrt{n}}\right) = 0.95 \Rightarrow$$

[4]

$$\Pr\left(\overline{X} + H*\frac{s}{\sqrt{n}} > \mu > \overline{X} - H*\frac{s}{\sqrt{n}}\right) = 0.95$$

We can say that there is a 95% chance that the mean $\mu$ will lie in the range of $\overline{X} \pm H*\frac{s}{\sqrt{n}}$.

We can replace H by $t_{1-\alpha/2,\, n-1}$, where $\alpha$ is referred to as rejection region used in hypothesis.

e.g. Let's assume that $\overline{X} = 68$ (student weight), s = 10 and n = 36 from a example of students' weight collected.

$$68 \pm 2.03*\frac{10}{\sqrt{36}} \cong 68 \pm 3.38$$

95% chance that mean $\mu$ will fall in the range (64.62, 71.38).

e.g. In a random sample of 64 law firms, legal charges per hour are found to have a mean of \$40 with a standard deviation 5. Obtain a 99% CI for the average legal charge per hour in the law profession as a whole.

[2] If variance is known, we can use the standard normal distribution (Z).

[3] H varies with the size of confidence interval as well as the distribution used (Z or t).

[4] $a < X < b \Rightarrow -a > -X > -b$

## C. Hypothesis Testing

Def: A statistical hypothesis is a conjecture about the sampling distribution of a random variable. When a hypothesis completely specifies the distribution, it is called a simple hypothesis; if it is not, it is referred to as a composite hypothesis.

Neyman and Pearson introduce their hypothesis testing approach by specifying the null hypothesis and alternative hypothesis. The former is denoted by $H_0$ and the latter $H_1$ or $H_A$. The null statement is a status quo and our goal is show that whether the evidence from the data is "strong" enough to reject the null hypothesis.

e.g.

$H_0$: $\mu = 42{,}000$
$H_1$: $\mu = 45{,}000$

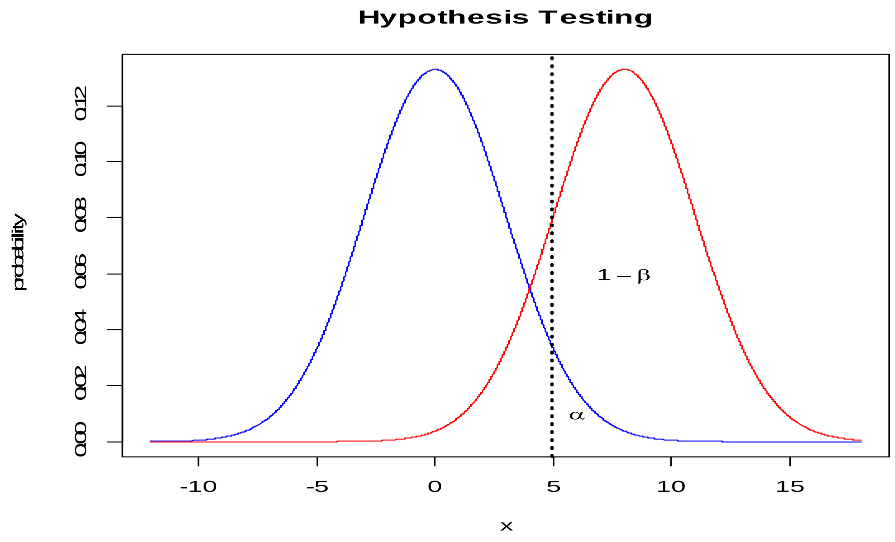
$H_0$: $\mu \leq 42{,}000$
$H_1$: $\mu > 42{,}000$

Type I and Type II Error

| | $H_0$ is true | $H_0$ is false |
|---|---|---|
| Accept $H_0$ | | Type II ($\beta$) |
| Reject $H_0$ | Type I ($\alpha$) | |

Power = 1 - $\beta$

Power: the ability to detect a false null hypothesis.
In practice we like our test to have high power given the reject region.



Note: $\alpha$ = 0.05 is the rejection region.

**One-tail Test**

Let's use the previous numerical example from page 5. There we assume that $\overline{X}$ = 68, s = 10 and n = 36. This example is about policy effectiveness. The tax is levied on sweet drinks in order to help children reduce their weight (childhood obesity is a popular research topic in recent years). We are interested in examining whether the average weight is "actually" decreased the sweet drinks tax?

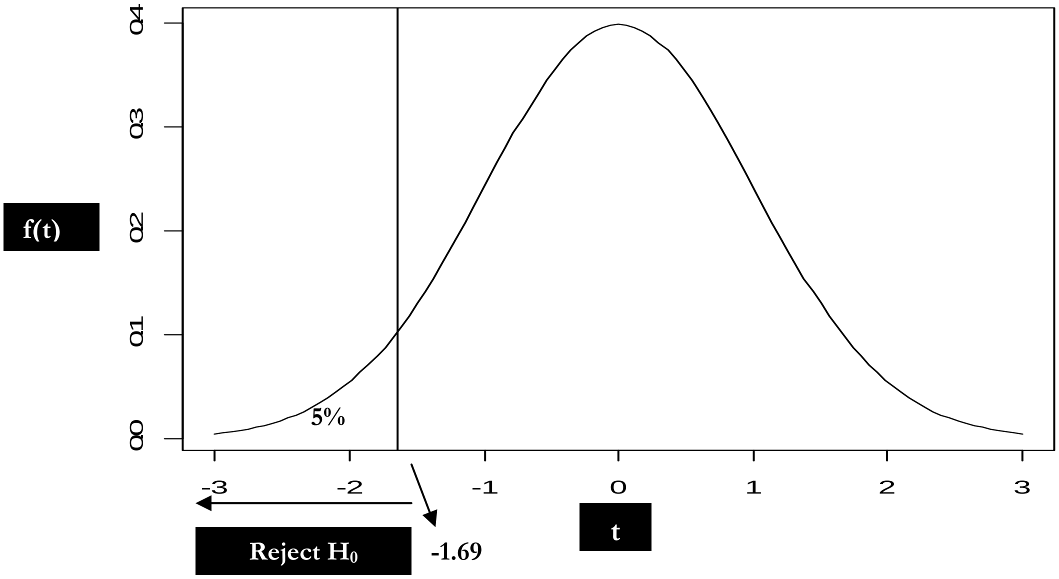
$H_0$: $\mu = 70$ … Null Hypothesis
$H_A$: $\mu < 70$ … Alternative Hypothesis

If the null hypothesis is true, then TS (test statistic) has a t distribution and can be represented as follows:

$$\text{TS} = \frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1} \Rightarrow \frac{\overline{X}-70}{s/\sqrt{n}} = \frac{68-70}{10/\sqrt{36}} = -1.2 \text{ … We can't not reject that } \mu = 70.$$

Conclusion: the new tax on sweet food may not be effective to reduce the students' weight.

Let's choose a level of significance that equals 5%[5]; the area under the curve and on the left of the value -1.69. Given the null is true, the chance that TS would fall in the level of significance is only 5%. Meaning, it is unlikely to happen. Therefore, the reasonable judgment is to reject the null hypothesis and accept alternative hypothesis. Note that there is a 5% chance that the null can be true. If this is the case, we may make the type I error[6].



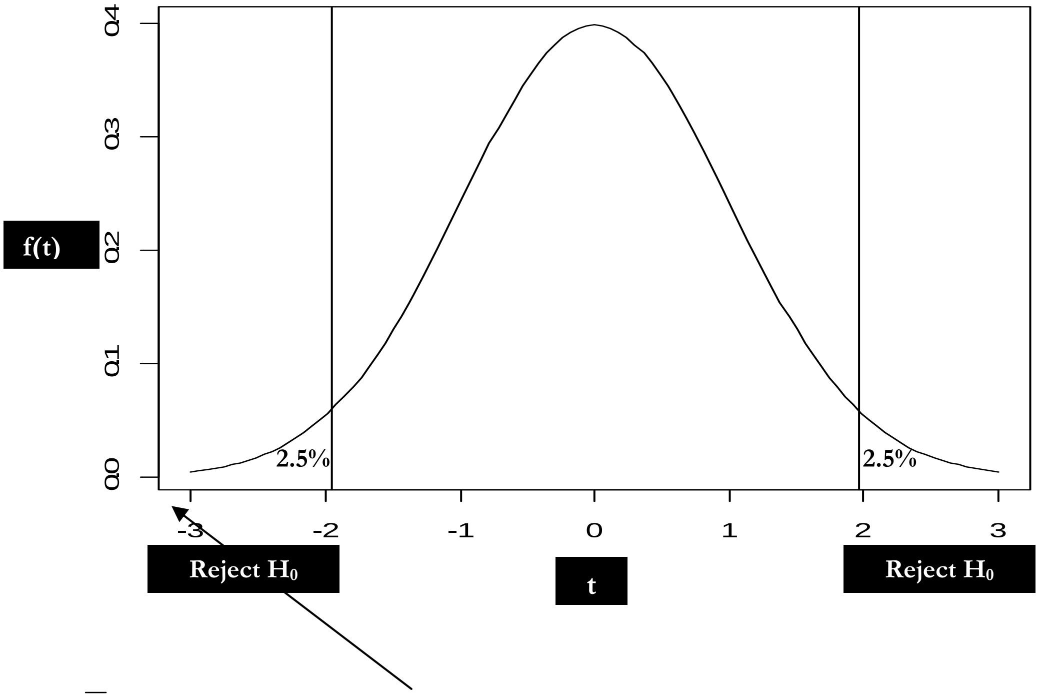
[5] It can also be 1% or 10% depends on researcher's choices.

[6] The Type II error refers to the situation where we accept a false null hypothesis. To reduce Type I error would result in an increase of Type II error. In practice we either reject the null hypothesis or reserve the judgment about it. By doing this we never accept null hypothesis and therefore we **cannot** make Type II error.

### Two-tail Test

e.g. The mean lifetime of a random sample of 80 light bulbs produced by a factory is found to be 1460 hours, with a standard deviation s = 110 hours. If $\mu$ is the mean lifetime of all the light bulbs produced by the factory, test the hypothesis $\mu = 1500$ against the alternative hypothesis that $\mu \neq 1500$, using the level of significance 0.05.

$H_0$: $\mu = 1500$ … Null Hypothesis
$H_A$: $\mu \neq 1500$ … Alternative Hypothesis



TS = $\frac{\overline{X}-\mu}{s/\sqrt{n}} = \frac{1460-1500}{110/\sqrt{80}}$ = -3.25 … We reject that $\mu = 1500$.

***The other method to conduct hypothesis testing is to use P-value. P-value is defined as the lowest significance level at which the null hypothesis can be rejected. Let's use the one-tail test from page 7. Since the TS is -1.2, we can find the P-value is 0.119 or 11.9%. Graphically, it is an area on the left of TS (-1.2) under the t curve.