# ECO220: Quantitative Methods in Economics
# Lecture Notes (0201)

Tianyu Du, Instructor: Victor Yu

July 10, 2018

## Contents

**1 Lecture 1 May. 08 2018** **3**
- 1.1 Notations . . . 3
- 1.2 From Sample to Population . . . 3

**2 Lecture 2 May. 09 2018** **3**
- 2.1 What is Statistics? . . . 3
- 2.2 Data . . . 3
- 2.3 Descriptive Statistics - Graphs . . . 4
- 2.4 Descriptive Statistics - Numerical measures . . . 4
  - 2.4.1 Measures of centre . . . 4

**3 Lecture 3 May. 15 2018** **4**
- 3.1 Measures of Variation(Spread) . . . 4

**4 Lecture 4 May. 16 2018** **5**
- 4.1 Covariance and Correlation: on Populations . . . 5
- 4.2 Covariance and Correlation: on Samples . . . 5
- 4.3 Interpretations . . . 6
  - 4.3.1 Interpreting covariance . . . 6
  - 4.3.2 Interpreting correlation coefficient . . . 6

**5 Lecture 5 May. 22 2018** **6**
- 5.1 Introduction to Simple Regression . . . 6
- 5.2 Relationship between $b_1$ and $r$ . . . 7
- 5.3 Analysis of Variance (ANOVA) . . . 7

**6 Lecture 6 May. 23 2018** **8**
- 6.1 OLS, continued. . . . 8
- 6.2 Sample space, Event and Probability . . . 8
- 6.3 Some Rules of Probability . . . 8

**7 Lecture 7 May. 29 2018** **9**
7.1 Conditional Probability . . . 9
7.2 Independent Event . . . 9

**8 Lecture 8 May. 30 2018** **9**
8.1 Bayes Theorem . . . 9
8.2 Random Variable and Prob. Distributions . . . 9
8.3 Expected Values . . . 10

**9 Lecture 9 June. 5 2018** **10**
9.1 Expected Value of a Random Variable . . . 10
9.2 Laws of Expectation . . . 10
9.3 Binomial Distribution . . . 10

**10 Lecture 10 June. 6 2018** **11**
10.1 Uniform Distribution . . . 11
10.2 Normal Distribution . . . 11

**11 Lecture 11 June. 12 2018** **11**
11.1 Applying normal distribution . . . 11
11.2 Normal Approximation to Binomial . . . 11

**12 Lecture 12 June. 13 2018** **12**
12.1 Sampling Distributions . . . 12
12.2 Sampling distribution of $\overline{X}$, the sample mean . . . 13

**13 Lecture 13 Jun. 19 2018** **13**
13.1 Confidence Interval . . . 13
13.2 Sample Size Required . . . 13

**14 Lecture 14 Jul. 3 2018** **14**
14.1 Confidence Interval for Population Proportion . . . 14
14.2 Two populations . . . 14
14.3 Chapter 12. Hypothesis Testing in Population Proportion . . . 15

**15 Lecture 15 Jul. 4 2018** **16**
15.1 Recall: Concepts of Hypothesis Testing . . . 16
15.2 Hypothesis Testing . . . 16
15.3 Hypothesis Testing in Population Proportion . . . 17
15.3.1 Finding Critical Value c . . . 17

**16 Lecture 16 Jul. 10 2018** **18**
16.1 Concepts in Hypothesis Testing . . . 18
16.2 Hypothesis on $p$, Population Proportion . . . 18
16.3 Determining the Critical Value $c$ . . . 18

# 1 Lecture 1 May. 08 2018

## 1.1 Notations

| Variable | Population | Sample |
|---|---|---|
| Size | $N$ | $n$ |
| Mean | $\mu$ | $\overline{x}$ |
| Std | $\sigma$ | $s$ |

## 1.2 From Sample to Population

Let $p$ denote the percentage of qualified people in <u>population</u> and let $\hat{p}$ denote the percentage of qualified people in <u>sample</u>. Then, $p$ has an <u>unknown</u> value and the value $\hat{p}$ can be calculated from sample data. We say *$\hat{p}$ is an **estimator** for $p$*, and the value of $p$ is still unknown and can only be estimated.

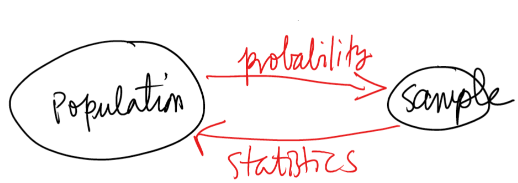
$p$ is a **fixed value** (i.e. $p$ is fixed once population is fixed, we can measure the exact and certain value of $p$ if we traverse the whole population). But $\hat{p}$ will change from sample to sample. We call $\hat{p}$ an **estimator** (or **sample statistic**). The value of sample statistic will change from sample to sample. And, therefore, we call $\hat{p}$ a **random value**.

# 2 Lecture 2 May. 09 2018

## 2.1 What is Statistics?

$$\text{Statistics}\begin{cases}\text{Descriptive Statistics}\begin{cases}\text{Graphs}\\ \text{Numerical measures}\end{cases}\\ \text{Inferential Statistics} \quad \textit{Draw conclusions in a population based on sample data.}\end{cases}$$

*Inferential Statistics involves uncertainties. To deal with the uncertainties, we need **probability***




## 2.2 Data

$$\text{Data}\begin{cases}\text{Quantitative data}\begin{cases}\text{Discrete}\\ \text{Continuous}\end{cases}\\ \text{Qualitative data (Categorical data)}\end{cases}$$

### 2.3 Descriptive Statistics - Graphs

### 2.4 Descriptive Statistics - Numerical measures

#### 2.4.1 Measures of centre

**Mean** Let $\{x_1, \ldots, x_N\}$ be measurements for the population with size $N$. The population mean is denoted by $\mu$ and defined as

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Let $\{x_1, \ldots, x_n\}$ be measurements for the sample of size $n$. The sample mean is denoted by $\overline{x}$ and defined as

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

**Note** *The mean is sensitive to extreme values.*

**Median** is the value in the middle when all data are put in order of magnitude. (For data with even size, median is defined as the average of the two values in the middle.)

**Mode** is value(s) with highest frequency.

**Percentiles** the $k^{th}$ percentile is a number such that $k\%$ of data fall below this number.

## 3 Lecture 3 May. 15 2018

### 3.1 Measures of Variation(Spread)

**Variance and Standard Derivation** Let $\{x_1, \ldots, x_N\}$ denote the population with size $N$ and let $\{x_1, \ldots, x_n\}$ denote the sample with size $n$. Then

| Measures | Population | Sample |
|---|---|---|
| Size | $N$ | $n$ |
| Mean | $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$ | $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ |
| Variance | $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$ | $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$ |
| Std | $\sigma = +\sqrt{\sigma^2}$ | $s = +\sqrt{s^2}$ |

**Note** When calculate the sample variance, use $n - 1$ as denominator.

**Note** mathematically,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$$

**Range** is defined as the difference between the largest value and the smallest value.

# 4 Lecture 4 May. 16 2018

## 4.1 Covariance and Correlation: on Populations

Consider two sets of data (population) with size $N$, denoted as $\{x_1, \dots, x_N\}$ and $\{y_1, \dots, y_N\}$, where $x$ and $y$ measure the age and income of observation, respectively.

**Denote** $\mu_x :=$ mean of $x$, $\mu_y :=$ mean of $y$
$\sigma_x :=$ std dev of $x$ and $\sigma_y :=$ std dev of $y$. *When $x$ changes, does $y$ change?*

**Covariance** defined covariance between two datasets, $x$ and $y$ as,

$$Cov(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

**Correlation coefficient** the correlation coefficient $\rho$ between datasets $x$ and $y$ is defined as

$$\rho = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

## 4.2 Covariance and Correlation: on Samples

When $N$ is too large, we select a sample of size $n$.

Let $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_n\}$ denote the selected samples with size $n$, $\overline{x}, \overline{y}$ denote the sample means, and $s_x, s_y$ denote the sample std dev.

**Covariance** between two sample is defined as

$$Cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

**Correlation coefficient** the sample correlation $r$ is defined as

$$r = \frac{Cov(x, y)}{s_x s_y}$$

## 4.3 Interpretations

### 4.3.1 Interpreting covariance

**Example** Consider samples $x$ and $y$ with

$$Cov(x, y) = -25.31$$

The negative sign means $x$ and $y$ have a negative linear relationship. As $x$ increases, $y$ tends to decrease. The magnitude 25.31 has **no** meaning.

### 4.3.2 Interpreting correlation coefficient

**Example** Consider samples $x$ and $y$ with

$$r = -0.94$$

The negative sign means $x$ and $y$ have negative linear relationship. As $x$ increases, $y$ decreases. The magnitude 0.94 means the linear relationship is strong. When $r$ is close to 1 or -1, the string line relation is strong, when $r$ is close to 0, the relation is weak.

**Note** $\rho \in [-1, 1]$ and $r \in [-1, 1]$

# 5 Lecture 5 May. 22 2018

## 5.1 Introduction to Simple Regression

Let the linear estimator to be $\hat{y} = b_0 + b_1 x$ and let $y_i$ denote the actual value at $x_i$, $\hat{y}$ is the estimated $y$ value at $x_i$. Then, $e_i := y_i - \hat{y_i}$ is the error of y value at $x_i$ (a.k.a. **residual**).

**Note** notice that $\sum_{i=1}^{n} e_i \equiv 0$.

**SSE** **Sum of Squared Error**(SSE) as

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

**OLS** Minimize $SSE$ with respect to $b_0$ and $b_1$, we have the FOC as

$$\begin{cases} \frac{\partial SSE}{\partial b_0} = 0 \\ \frac{\partial SSE}{\partial b_1} = 0 \end{cases}$$

By solving the first order conditions, we have

$$\begin{cases} b_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} \\ b_0 = \overline{y} - b_1 \overline{x} \end{cases}$$

The above method to find $b_0$ and $b_1$ is called the <u>method of least square</u>, or <u>method of Ordinary Least Square (OLS)</u>.

## 5.2 Relationship between $b_1$ and $r$

$$b_1 = \frac{Cov(x,y)}{Var(x)} = \frac{Cov(x,y)}{std(x)std(y)} \frac{std(y)}{std(x)} = r\frac{s_y}{s_x}$$

## 5.3 Analysis of Variance (ANOVA)

Let $y_i$ denote the actual $y$ value at $x_i$ and $\hat{y}_i$ denote the estimated $y$ value at $x_i$.

**Definition**

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_1)^2$$

**Notice** $SST = SSR + SSE$

**Anova Table**

| | SS | df | MS | F |
|---|---|---|---|---|
| Regression | SSR | 1 | MSR | MSR/MSE |
| Error(Residual) | SSE | MSE | $n-2$ | |
| Total | SST | $n-1$ | | |

where MS stands for **mean square** and is defined as

$$MS = \frac{SS}{df}$$

$$MSR = \frac{SSR}{1}$$

$$MSE = \frac{SSE}{n-2}$$

# 6 Lecture 6 May. 23 2018

## 6.1 OLS, continued.

**R-square** **coefficient of determination** is defined as

$$R^2 = \frac{SSR}{SST}$$

and notice that $R^2 \in [0, 1]$ and can be interpreted as **% of variation in** $y$ **explained by** $x$ **(via the linear model)**

**Note** in ECO220, we use $R^2$ or $r^2$ to represent the same thing.

## 6.2 Sample space, Event and Probability

**Experiment** an experiment is a process that creates two or more outcomes.

**Random Experiment** a random experiment is an experiment such that the outputs *cannot* be determined with certainty before the end of the experiment.

**Sample Space** a sample space is the set of all possible outcomes in a random experiment.

**Event** an event is a subset of a sample space.

**Prob** Let $S$ be the sample space, let $E$ be an event, then the **probability of** $E$, $P(E)$ is defined as

$$P(E) = \text{probability of } E = \frac{\text{Number of outcomes in } E}{\text{Number of outcomes in } S}$$

*assuming that each outcome in $S$ has equal likelihood to be chosen into $E$.*

## 6.3 Some Rules of Probability

Let $E$ be an event in sample space $S$, then

- $P(E) \in [0, 1]$.
- $P(S) = 1$.
- Let $E^c$ denote the **complementary** of $E$, then $P(E^c) = 1 - P(E)$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (Addition Rule)

# 7 Lecture 7 May. 29 2018

**Mutually Exclusive Event** If $A \cap B = \emptyset$, we say events $A$ and $B$ are **mutually exclusive/disjoint**. Then, if $A, B$ are disjoint, we have

$$P(A \cup B) = P(A) + P(B)$$

## 7.1 Conditional Probability

**Conditional Prob** In general, if $A$ and $B$ are events in sample space $S$, the <u>conditional probability of $A$ given $B$</u> is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

**Multiplication rule**

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

## 7.2 Independent Event

**Independent Event** We say two events, $A$ and $B$ are **independent** if any of the following is true. (those definitions below are equivalent.)

- $P(A|B) = P(A)$ or
- $P(B|A) = P(B)$ or
- $P(A \cap B) = P(A)P(B)$

# 8 Lecture 8 May. 30 2018

## 8.1 Bayes Theorem

Let $A$ and $B$ be two events. Then,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

*Proven by definition of conditional probability.*

## 8.2 Random Variable and Prob. Distributions

**Prob. distribution**

**Cumulative Prob. distribution**

## 8.3 Expected Values

**Expected Value** Let $X$ be a random variable with probability distribution $P(X)$. Then defined the expected value of $X$, $\mathbb{E}(X)$ as

$$\mu = \mathbb{E}(X) = \sum_x xP(X = x)$$

**Variance of Random Variable** For random variable $X$, we have

$$\sigma^2 = Var(X) = \sum_x (x - \mu)^2 P(X = x) = \mathbb{E}(X - \mu)^2$$

# 9 Lecture 9 June. 5 2018

## 9.1 Expected Value of a Random Variable

**Mean** $\mu = \mathbb{E}(X) = \sum_x xP(X = x)$.

**Variance** $\sigma^2 = \mathbb{E}(x - \mu)^2 = \mathbb{E}(X^2) - \mu^2$.

## 9.2 Laws of Expectation

In general, let $X$ be a random variable, and let $a, c \in \mathbb{R}$, then

$$\mathbb{E}(aX + c) = a\mathbb{E}(X) + c$$

$$Var(aX + c) = Var(aX) = a^2 Var(X)$$

Let $X$ and $Y$ be random variables, and let $a, b, c \in \mathbb{R}$, then

$$\mathbb{E}(aX + bY + c) = a\mathbb{E}(X) + b\mathbb{E}(Y) + c$$

$$Var(aX + bY + c) = Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2ab\ Cov(X, Y)$$

**Note** if $X$ and $Y$ are independent, then $\rho = Cov(X, Y) = 0$ and

$$Var(aX + bY + c) = a^2 Var(X) + b^2 Var(Y)$$

## 9.3 Binomial Distribution

In general, let $n$ be the number of independent trails and $p = P(\#success)$. Let $X$ be a random variable which is the number of successes in $n$ trails, we have

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x},\ \text{for } x = \{0, 1, 2, \dots, n\}$$

$$\mu = \mathbb{E}(X) = np$$

$$\sigma^2 = Var(X) = npq,\ q = 1 - p$$

# 10 Lecture 10 June. 6 2018

## 10.1 Uniform Distribution

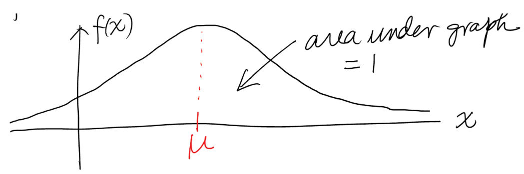
Let $X$ be uniform from $a$ to $b$. $f(x) = \frac{1}{b-a}, a \leq x \leq b$

$$\mu = \mathbb{E}(X) = \int_a^b xf(x)dx = \frac{a+b}{2}$$

$$\sigma^2 = Var(X) = \mathbb{E}(X^2) - \mu^2$$

## 10.2 Normal Distribution

Let $X$ be a continuous random variable, satisfying $-\infty < x < \infty$. The mean of $X$ is $\mu$ and the variance of $X$ is $\sigma^2$. The graph of $X$ is The graph is




symmetric at $\mu$ and the variance $\sigma^2$ determines the shape(spread) of $X$. We say $X$ follows a normal distribution with mean $\mu$ and variance $\sigma^2$. And denote as

$$X \sim N(\mu, \sigma^2)$$

**Standard Normal Distribution** A standard normal distribution is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. Denote the standard normal distribution as

$$Z \sim N(0, 1)$$

# 11 Lecture 11 June. 12 2018

## 11.1 Applying normal distribution

**Theorem** Let $X \sim N(\mu, \sigma^2)$, then

$$z = \frac{x - \mu}{\sigma} \sim N(0, 1)$$

## 11.2 Normal Approximation to Binomial

Consider a random variable $X \sim B(n, p)$, then we can approximate the binomial with a normal distribution $X \approx N(np, npq)$.

# 12 Lecture 12 June. 13 2018

## 12.1 Sampling Distributions

Consider population with size $N$ has $p$ as percentage of *success* (qualified) and sample with size $n$ with $\hat{p} = \frac{x}{n}$ as percentage of *success*.

$p$ is a parameter which has a fixed value. In real life, the value of $p$ is usually unknown. $\hat{p}$ is a sample statistic, which does not have fixed value (random variable, value of $\hat{p}$ vary from sample to sample). Also, $\mu$ and $\sigma$ are parameters, which are fixed but usually unknown. $\overline{x}$ is a sample statistic, and is random.

Suppose we know $p$ for population, then we can conclude about random variables from a random sample,

1. $\mathbb{E}(\hat{p}) = p$.
2. $Var(\hat{p}) = \frac{pq}{n},\ q = 1 - p$
3. When sample size $n$ is large, the distribution of $\hat{p}$ is <u>approximately normal</u> (**Central Limit Theorem in proportion**) [1]

That's

$$\hat{p} \approx\sim N(p, \frac{pq}{n}),\ \text{when } n \text{ is large.}$$

**Example** Given $p_{success} = 0.3$ for the whole population and find the probability that at least 320 *success* found in a sample of size $n = 1000$. i.e. Let $X$ denote the number of success in sample with $n = 1000$, find $P(X \geq 320)$.

**Method 1** Use Central Limit Theorem, check $np = 300 \geq 10 \wedge nq = 700 \geq 10$, thus $n$ is *large*. And approximate $\hat{p}$ of sample as

$$\hat{p} \sim N(p, \frac{pq}{n})$$

*Soln.*

$$P(X \geq 320) = P(\hat{p} \geq 0.32)$$
$$= P(\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \geq \frac{0.32 - 0.3}{\sqrt{\frac{0.3*0.7}{1000}}}) = P(z \geq \frac{0.02}{\sqrt{\frac{0.21}{1000}}})$$

Find $z$ in z-table

■

[1] As a rule of thumb, $n$ is considered to be large when $np \geq 10 \wedge nq \geq 10$.

**Method 2** Use Normal Approximation to Binomial. $p = 0.3$ and $n = 1000$.

$$X \approx Y \sim N(300, 210)$$

*Soln.*

$$\begin{aligned} P(X \geq 320) &= P(Y > 319.5) \\ &= P(\frac{Y-\mu}{\sigma} > \frac{319.5 - 300}{\sqrt{210}}) \\ &= P(z > 1.35) \text{ find in z table} \end{aligned}$$

■

**Note** *methods 1 and 2 do* ***not*** *give exactly same answer, but the answers should be close.*

### 12.2 Sampling distribution of $\overline{X}$, the sample mean

1. $\mathbb{E}(\overline{X}) = \mu$.
2. $Var(\overline{X}) = \frac{\sigma^2}{n}$.
3. When $n$ is large, the distribution of $\overline{X}$ is approximately normal. (**Central Limit Theorem in Mean**).
4. When population is normal, the distribution of $\overline{X}$ is exactly normal, regardless of the sample size $n$.

Putting together,

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}), \text{ when } n \text{ is large.}$$

## 13 Lecture 13 Jun. 19 2018

### 13.1 Confidence Interval

To find $100(1-\alpha)\%$ confidence interval for $p$ estimated from $\hat{p}$ is

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

### 13.2 Sample Size Required

When we specify the confidence level $1-\alpha$, and the margin of error, the required sample size is

$$n = \frac{z_{\frac{\alpha}{2}}^2}{(ME)^2} pq$$

If $p$ can be estimated from previous surveys, use it to find $n$. Else, use $p = 0.5$ to find $n$.

## 14 Lecture 14 Jul. 3 2018

### 14.1 Confidence Interval for Population Proportion

**Point estimator** for $p$ is $\hat{p}$, confidence interval for $p$ is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}},\ \textit{with large } n$$

$n$ is considered as *large* iff $np \geq 10 \wedge nq \geq 10$. $\sqrt{\frac{\hat{p}\hat{q}}{n}}$ is the **standard error/deviation** of estimation. And $z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$ is the **margin of error**.

### 14.2 Two populations

$p_1, p_2$ denote the qualification percentages for population 1 and 2. And $\hat{p_1}, \hat{p_2}$ denote the qualification percentages in samples with sample sizes $n_1, n_2$ from population 1 and 2.

**Point estimator** To estimate $p_1$ to $p_2$, we estimate $p_1 - p_2$. The point estimator for $p_1 - p_2$ is $\hat{p_1} - \hat{p_2}$.

**Interval estimator** The interval estimation for $p_1 - p_2$ is

$$PointEstimator \pm z_{\alpha/2} \times Std(PointEstimator)$$

$$\hat{p_1} - \hat{p_2} \pm z_{\alpha/2} \times Std(\hat{p_1} - \hat{p_2})$$

To find $Std(\hat{p_1} - \hat{p_2})$, by *law of expectation*

$$Var(aX + bY) = Var(aX) + Var(bY) + 2abCov(X, Y)$$

We select two independent samples of size $n_1$ and $n_2$ from populations, therefore

$$V(\hat{p_1} - \hat{p_2}) = V(\hat{p_1}) + V(\hat{p_2})$$

When $n_1$ and $n_2$ are <u>large</u>, by *central limit theorem*,

$$\hat{p_1} \sim N(p_1, \frac{p_1 q_1}{n_1}), \quad \hat{p_2} \sim N(p_2, \frac{p_2 q_2}{n_2})$$

Then, for two *independent* samples, $\hat{p_1}$ and $\hat{p_2}$ are independent,

$$V(\hat{p_1} - \hat{p_2}) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$$

But we do not know $p_1$ and $p_2$, we cannot calculate $V(\hat{p_1} - \hat{p_2})$ directly from above equation. We estimate $p_1$ and $p_2$ by $\hat{p_1}$ and $\hat{p_2}$
Therefore the **estimated variance** is

$$(Estimated)V(\hat{p_1} - \hat{p_2}) = \frac{\hat{p_1}\hat{q_1}}{n_1} + \frac{\hat{p_2}\hat{q_2}}{n_2}$$

$$(Estimated)Std(\hat{p_1} - \hat{p_2}) = \sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1} + \frac{\hat{p_2}\hat{q_2}}{n_2}}$$

**Result** confidence interval for $\hat{p_1} - \hat{p_2}$

$$C.I._\alpha = \hat{p_1} - \hat{p_2} \pm z_{\alpha/2}\sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1} + \frac{\hat{p_2}\hat{q_2}}{n_2}}$$

**Example** Compare the percentage of people going to casino in Ontario and Manitoba.

| Var | Ontario | Manitoba |
|---|---|---|
| Actual | $p_1$ | $p_2$ |
| Sample size | $n_1 = 4151$ | $n_2 = 389$ |
| Point estimator | $\hat{p_1} = 66.5\%$ | $\hat{p_2} = 75.2\%$ |

**Point estimator** for $p_1 - p_2$ is $\hat{p_1} - \hat{p_2} = -0.087$
the 95% C.I. for $p_1 - p_2$ is

$$-.0087 \pm (z_{.025} = 1.96) * \sqrt{\frac{.665 * .335}{4151} + \frac{.752 * .248}{389}} = (-.132, -.042)$$

**Interpretation** 95% of the times, $p_1 - p_2$ falls between -.132 and -.042. We are 95% confident that $p_1 - p_2$ is between -.132 and -.042.
**Remark** Is there a significant difference between the % going to casinos between Ontario and Manitoba? Since the 95% C.I. for $p_1 - p_2$ does not contain 0, we conclude that there **is** a significant difference between % in two population.
**Remark** You can also use estimate the $p_2 - p_1$ and the 95% C.I. would be $(.042, .132)$.

## 14.3 Chapter 12. Hypothesis Testing in Population Proportion

$$\text{Statistical Inference}\begin{cases}\text{Estimation}\begin{cases}\text{Point Estimation}\\ \text{Interval Estimation}\end{cases}\\ \text{Hypothesis Testing}\end{cases}$$

$H_0$: the **null hypothesis**. $H_1$: the **alternative hypothesis**.

**Reality**

$$\begin{cases}H_0 \text{ Person not murder}\begin{cases}Guilty \mid H_0 \ \textbf{Type I Error}\\ NotGuilty \mid H_0 \ \textbf{No error}\end{cases}\\ H_1 \text{ Person is murderer}\begin{cases}Guilty \mid H_1 \ \textbf{No error}\\ NotGuilty \mid H_1 \ \textbf{Type II Error}\end{cases}\end{cases}$$

When the court concludes "Guilty" and $H_0$ is true, type I error occurs and denote the probability of type I error as $\alpha$

$$\alpha = P(\text{Reject } H_0|H_0)P(\text{Type I Error})$$

And in this case, Type II Error will not occur.

Let $\beta$ denote the probability for type II error to occur.

$$\beta = P(\text{Reject } H_1/\text{Fail to reject } H_0|H_1) = P(\text{Type II Error})$$

**Remark** $\beta$ becomes large as $\alpha$ becomes small.

**Conclusion** Therefore, in *hypothesis testing*, we have two hypotheses, $H_0$ and $H_1$. Based on sample results (evidence), we either *reject* $H_0$ or *do not reject* $H_0$

# 15 Lecture 15 Jul. 4 2018

## 15.1 Recall: Concepts of Hypothesis Testing

**Concepts** $H_0$ **null hypothesis** and $H_1$ **alternative hypothesis**

**Case I** Rejecting $H_0$ while $H_0$ is true. <u>Failed to accept null hypothesis $H_0$</u>.

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ True}) = \textbf{Significance Level}$$

*In real life, we want the **significance level** to be as low as possible.* And note that significance level is probability for <u>type I error</u> to occur when $H_0$ is true.

**Case II** Accepting $H_0$ with $H_1$ is true. <u>Fail to accept alternative hypothesis</u>. $\beta$ is the probability of <u>type II error</u> while $H_1$ is true.

$$\beta = P(\text{Accept } H_0 \mid H_1 \text{ True})$$

**Remark** In hypothesis testing, we wish both $\alpha$ and $\beta$ to be small. However, by setting the rule of decision making and lowering $\alpha$, $\beta$ goes higher.

## 15.2 Hypothesis Testing

**Steps**

1. Set up null and alternative hypotheses $H_0$ and $H_1$.
2. Setup a decision rule.
3. Test statistic.
4. Conclusion.

## 15.3 Hypothesis Testing in Population Proportion

**Data** Population proportion $p$. Select sample with sample size $n$ and sample proportion $\hat{p}$.

$H_0$ null hypothesis on $p$ and $H_1$ as the alternative hypothesis.

**Example** Government claims that more than 50% of Canadian are in favour of a policy.
**Step 1** setup hypotheses $H_0 := p < 0.5$ and $H_1 := p > 0.5$
**Step 2** setup a decision rule If $\hat{p} < c$ we accept $H_0$. If $\hat{p} > c$ we reject $H_0$.
Consider $c = 0.55$. Then in a random sample of $n = 200, \hat{p} = \%$ in favour from sample. Decision rule is

$$\begin{cases} \text{Accept } H_0 : p \leq 0.5 \text{ if } \hat{p} < 0.55 \\ \text{Accept } H_1 : p > 0.5 \text{ if } \hat{p} \geq 0.55 \end{cases}$$

**Step 3** test statistic. In a sample of $n = 200, \hat{p} = 0.58$.
**Step 4** Conclusion: Reject $H_0$.

### 15.3.1 Finding Critical Value c

Finding $\alpha$

$$\alpha = P(\text{Type I Error}) = P(\hat{p} \geq 0.55 | p \leq 0.5)$$

By central limit theorem, when $n$ is large,

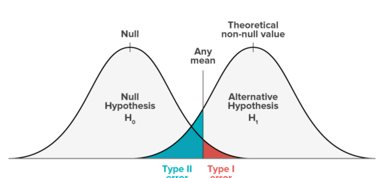
$$\hat{p} \sim N(np, \frac{pq}{n})$$

Then normalized data,

$$\alpha = p(z \geq \frac{0.05\sqrt{200}}{0.5}) = 0.0793 \approx 8\%$$

Therefore, by setting the critical value $c = 0.55$, the probability of type I error, $\alpha = 0.08$.
Finding $\beta$

$$\beta = P(\hat{p} < 0.55 | p > 0.5)$$



# 16 Lecture 16 Jul. 10 2018

## 16.1 Concepts in Hypothesis Testing

1. Set up $H_0$ and $H_1$ hypotheses.
2. Set up a decision rule.
3. Test statistic (sample evidence).
4. Conclusion.

## 16.2 Hypothesis on $p$, Population Proportion

**Example** Government Claims that more than half of the Canadian support the idea of increasing [...]

$$H_0 : p \leq 0.5, \quad H_1 : p > 0.5$$

**Decision rule** we will select a large sample, $n = 100$. Let $\hat{p} = \%$ support in the sample. By choosing critical value $c$, we reject $H_0$ if $\hat{p} \geq c$ and do not reject $H_0$ if $\hat{p} < c$.

## 16.3 Determining the Critical Value $c$

**Cont'd** Since $n = 100$ is large, by central limit theorem,

$$\hat{p} \sim N(p, \frac{pq}{n})$$

**Remark** To determine a critical value $c$, we select an acceptable $\alpha$ value.

**Example** Refer to previous example.

$$H_0 : p \leq 0.5, \quad H_1 : p > 0.5$$

Want to choose $\alpha = 0.01$. We know that if $\hat{p} \geq c$ then reject $H_0$. then

$$\begin{aligned}
\alpha = 0.01 &= P(\text{Reject } H_0 | H_0 \text{ True}) \\
&= P(\hat{p} \geq c | p \leq 0.5) \\
&\text{(Normalizing data, since } p \text{ is unknown, use 0.5 for it.)} \\
&= P(\frac{\hat{p} - p}{\sqrt{pq/n}} \geq \frac{c - 0.5}{\sqrt{(0.5)(0.5)/100}}) \\
&= P(z \geq \frac{c - 0.5}{\sqrt{(0.5)(0.5)/100}}) \\
&\implies \frac{c - 0.5}{\sqrt{(0.5)(0.5)/100}} = 2.33 \\
&\implies c = 0.5 + 2.33\sqrt{\frac{0.25}{100}} \\
&\implies c = 0.6165
\end{aligned}$$

■

Therefore, for decision rule $\alpha = 0.01$, corresponds to *Reject* $H_0$ *if* $c \geq 0.6165$. also can be stated as a normalized proportion: *Reject* $H_0$ *if* $z = \frac{\hat{p}-p}{\sqrt{pq/n}} \geq 2.33$

**Ways to Write Decision Rules**

1. By significance value $\alpha = 0.01$
2. By sample proportion If $\hat{p} \geq 0.6165$ then reject $H_0$, else accept $H_0$.
3. By normalized sample proportion If $z \geq 2.33$ reject $H_0$, else, accept $H_0$; where $z = \frac{\hat{p}-p}{\sqrt{pq/n}}$.

**Example** Manufacturer claims that less than 10% of the computer ship they manufactured are defective. Select a random sample with $n = 100$.
*Solution.*
Let $p = \%$ defective. Setting up the hypotheses:

$$H_0 := p \geq 0.1 \quad H_1 := p < 0.1$$

Setting up the decision rule: $\alpha = 0.05$. i.e. Reject $H_0$ if $\hat{p} \leq c$.

*Finding c:*

$$\begin{aligned}
\alpha = 0.05 &= P(\hat{p} \leq c | p \geq 0.1) \\
\implies 0.05 &= P(z = \frac{\hat{p} - p}{\sqrt{pq/n}} \leq \frac{c - 0.1}{\sqrt{0.09/100}}) \\
&\implies c = 0.05065
\end{aligned}$$

■

Therefore decision rule associated with $\alpha = 0.05$ is *if* $\hat{p} \leq 0.05065$ *then reject* $H_0$.
Sample Statistic: in the sample of $n = 100$, $\hat{p} = 0.06$. Therefore, since $\hat{p} > 0.005065$, accept $H_0$.

**Remark** Suppose we select the decision rule $\alpha$ less than 0.05, the conclusion from hypothesis testing remains unchanged. Suppose we select a larger $\alpha$ the conclusion might change, depending on how large $\alpha$ is.