AMS 572 Data Analysis I
Inference on one population mean μ

Pei-Fen Kuan

Applied Math and Stats, Stony Brook University

One population setting

Cross-sectional study: Collect data to test a hypothesis about the mean or median of X

- Take a random sample from the population.
- There are different types of sampling schemes. The simplest is the simple random sampling in which every subject in the population has equal chance to be selected.

Statistical inference on one population mean

Suppose we have a random sample of size n: $X_1, X_2, \ldots, X_n$ and we wish to draw inference about the population mean μ.

1. Point estimator
   - $\bar{X} = \frac{X_1+X_2+\ldots+X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$
   - Other estimators: median, mode, trimmed mean
2. Confidence Interval (C.I.)
3. Hypothesis Test
   - Example: μ is the height of adult US males
     H_0: $\leq 5^{'}6^{"}$ vs H_1: $\mu > 5^{'}6^{"}$

Recap: Normal Distribution

- Probability Density Function (p.d.f.)
  X follows normal distribution of mean μ and variance σ^2

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

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty, x \in R$$

  $P(a \leq X \leq b) = \int_a^b f(x)dx$ = area under the pdf curve bounded by a and b

- Cumulative Density Function (c.d.f.)

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

$$f(x) = [F(x)]' = \frac{d}{dx} F(x)$$

Recap: Normal Distribution

- Standard Normal Distribution

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

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

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

$$P(a \leq X \leq b) = P(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma})$$

$$= P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma})$$

Also known as **Z-score**

Proof: CDF Method

$$F_Z(z) = P(Z \leq z) = P(\tfrac{X-\mu}{\sigma} \leq z)$$
$$= P(X \leq \sigma \cdot z + \mu) = F_X(\sigma \cdot z + \mu)$$

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \frac{d}{dz} F_X(\sigma \cdot z + \mu) = f_X(\sigma \cdot z + \mu) \cdot \sigma$$

$$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\sigma \cdot z + \mu - \mu)^2}{2\sigma^2}} \cdot \sigma = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \text{ which is the p.d.f. for } N(0,1)$$

Proof: PDF Method/Jacobian Method

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$x = \sigma \cdot z + \mu$$

$$|J| = \frac{dx}{dz} = \frac{d}{dz}(\sigma \cdot z + \mu) = \sigma$$

$$f_z(z) = |J| \cdot f_x(x) = \sigma \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\sigma \cdot z+\mu-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

$$\therefore Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

Proof: MGF method

- Recall that
$$M_X(t) = E(e^{tX})$$
$= \int_{-\infty}^{\infty} e^{tx} f(x)dx$, if x is continuous
$= \sum_{x's} e^{tx} f(x)$, if x is discrete
- $M_Z(t) = e^{-\frac{\mu}{\sigma}t} \cdot M_X(\frac{1}{\sigma}t) = e^{-\frac{\mu}{\sigma}t} \cdot e^{\mu(\frac{1}{\sigma}t)+\frac{1}{2}\sigma^2(\frac{1}{\sigma}t)^2} = e^{\frac{1}{2}t^2}$
which is the m.g.f. for $N(0,1)$

Theorem: If two random variables have the same MGF, they have the same distribution.

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

$$M_X(t) \; = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

Eg) If the mgf of X is $e^{-7t+10t^2}$, then

Linear transformation of a normal random variable

- If $X \sim N(\mu, \sigma^2)$ and $Y = a \cdot X + b$, a, b are constants, then

$$Y \sim N(a\mu + b, a^2\,\sigma^2)$$

Example: Suppose the height of students of AMS 572 is normally distributed. Ten percent of the students are over 6.5 feet tall, while the variance is 0.390625 (or 0.625^2). What is the probability that the height of a student is in between 6 and 7 feet?

Distribution of sample mean $\bar{X}$

Theorem: Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with mean μ, variance σ^2. Then, the distribution of $\bar{X}$ is

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

Proof:

$$M_{\bar{X}}(t) = E(e^{t\bar{X}}) = E(e^{t\frac{\sum X_i}{n}}) = E(e^{\frac{\sum tX_i}{n}})$$

$$= E(\prod_{i=1}^{n} e^{\frac{t}{n}X_i}) = \prod_{i=1}^{n} \exp(\frac{\mu t}{n} + \frac{\sigma^2 t^2}{2n^2}) = \exp(\mu t + \frac{(\frac{\sigma^2}{n})t^2}{2}) \sim N(\mu, \frac{\sigma^2}{n})$$

Example (cont'd): If 10 students are chosen, what is the probability that the average height is in between 6 and 7 feet?

Inference for μ when σ^2 is known under normal distribution

Inference for μ when σ^2 is known under normal distribution

Setup:

- Assume that the distribution is normal
- Let $X_1, X_2, \ldots, X_n$ be a random sample for a normal distribution with mean μ and variance σ^2. That is, $X \overset{iid.}{\sim} N(\mu, \sigma^2), i = 1, ..., n$.
- Assume that σ^2 is known.

Inference for μ when σ^2 is known under normal distribution

1. Point estimator for μ

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

   Note that $E(\hat{\mu}) = E(\bar{X}) = \mu \Rightarrow \hat{\mu} = \bar{X}$ is an unbiased estimator of μ

   - $\bar{X}$ is also a maximum likelihood estimator (M.L.E.) and method of moment estimator of μ.

Inference for μ when σ^2 is known under normal distribution

2 Confidence Interval for μ:
Intuitive approach : $P(c_1 \leq \mu \leq c_2) = 0.95$

There ARE many ways to choose the c's.
We will show that for pivotal quantity with symmetric pdf, the symmetric CIs are the optimal, i.e., they have the shortest lengths for a given confidence level.

Pivotal Quantity (P.Q.) approach for deriving confidence interval

Definition: A pivotal quantity is a function of the sample and parameter of interest whose probability distribution does not depend on the unknown parameters.

Pivotal Quantity (P.Q.) approach for deriving confidence interval

Consider $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, where σ^2 is known

- Is $\bar{X}$ a pivotal quantity for μ?
- Function of $\bar{X}$ and μ : $\bar{X} - \mu \sim N(0, \frac{\sigma^2}{n})$
- Another function of $\bar{X}$ and μ : $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

Derivation for the symmetrical CI's for μ based on pivotal quantity Z

CI for μ, $0<\alpha<1$(e.g. α=0.05 $\Rightarrow$ 95% C.I.)

$$P(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}) = 1-\alpha$$

$$P(-Z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}) = 1-\alpha$$

$$P(-Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$P(\bar{X}-Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Thus the 100(1-α)% C.I. for μ is

Confidence Interval Interpretation

- If we draw 100 different random samples, on average $100(1-\alpha)\%$ of them will contain μ

  Sample size $n \longleftrightarrow$ the corresponding 100(1-α)% C.I.
  Sample 1, $\overline{x}$=5'7" $\longleftrightarrow$ [5'6",5'8"]
  Sample 2, $\overline{x}$=5'5" $\longleftrightarrow$ [5'4",5'6"]
  Sample 3, $\overline{x}$=5'8" $\longleftrightarrow$ [5'5",5'11"]
  …e.g.) 95% C.I., μ=5'7", 95% of all these CI's will cover μ

Confidence Interval Interpretation

- For the 100(1-α)% C.I. for μ is $[\bar{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}]$, the length of this CI is:

$$L_{sy} = 2 \cdot Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

- To decrease the length of confidence interval:
  - increase α, i.e., decrease confidence
  - increase sample size

Non-symmetric confidence interval

- Note that $P(-Z_{\alpha/3} \leq Z \leq Z_{2\alpha/3}) = 1 - \alpha$
- 100(1-α)% C.I. for μ

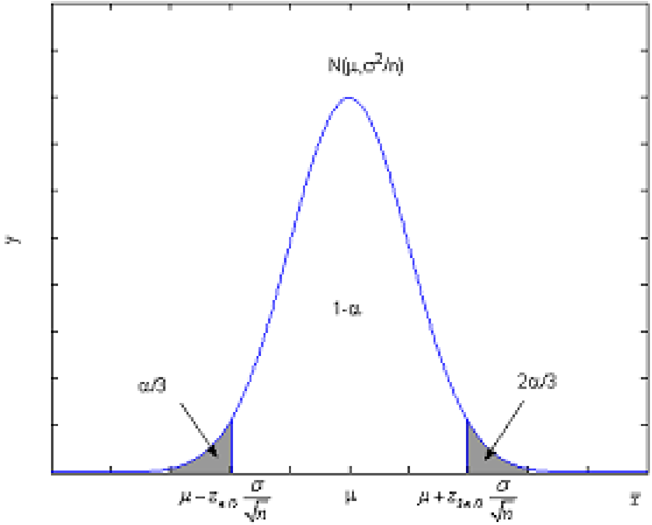
$$[\bar{X} - Z_{2\alpha/3} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/3} \cdot \frac{\sigma}{\sqrt{n}}]$$

  The length of this CI is:

$$L_{nsy} = (Z_{\alpha/3} + Z_{2\alpha/3}) \cdot \frac{\sigma}{\sqrt{n}}$$

- It can be shown that: $L_{sy} \leq L_{nsy}$
- Try a few numerical values for α, and see for yourself.

Non-symmetric confidence interval



Hypothesis testing

- Null hypothesis H_0: hypothesis to be tested
- Example of null hypothesis for folic acid study:
  The incidence of stroke will be the same in those taking folic acid supplements and those not taking folic acid supplements

Null and Alternative

- In a test of a hypothesis, we are testing whether some population parameter has a particular value or range
- For example,

  $$H_0 : \mu \leq \mu_0$$

  where μ_0 is a known constant
- The alternative hypothesis is complement of null hypothesis

  $$H_a : \mu > \mu_0$$

Null and Alternative

- A law suit analogy:
  Example: The O.J. Simpson trial
  H_0: OJ is innocent
  H_a: OJ is guilty

		The truth	
		H_0: OJ innocent	H_a OJ guilty
Jury's Decision	H_0	Right decision	Type II error
	H_a	Type I error	Right decision

Tests of Hypotheses: Seven Steps

1. Design study
2. Establish null hypothesis
3. Determine test statistic to be employed
4. Choose significance level α and establish critical region C_α. α is also $P(\text{Type I error})$.
5. Carry out study and collect data
6. Compute statistic from data
7. If statistic is in C_α, reject H_0

Types of hypothesis

<u>One-sided (or one-tailed) test:</u>

- $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$
  $H_0 : \mu \leq \mu_0$ vs $H_a : \mu > \mu_0$
- $H_0 : \mu = \mu_0$ vs $H_a : \mu < \mu_0$
  $H_0 : \mu \geq \mu_0$ vs $H_a : \mu < \mu_0$

<u>Two-sided (or two-tailed) test:</u>

- $H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$

Hypothesis test for $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$

<u>Setting</u>

- Data : $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2), \sigma^2$ is known
- The significance level α (e.g., $\alpha = 0.05$) is given.

Hypothesis test for $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$

Pivotal Quantity Method

1. Identify a pivotal quantity: Recall that

2. The **test statistic** is the pivotal quantity with the value of the parameter of interest under the null hypothesis

Hypothesis test for $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$

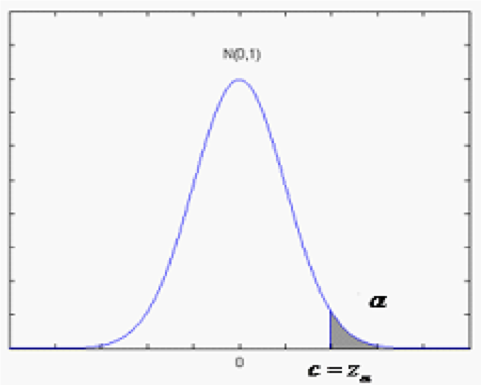
3 Derive the decision threshold for your test based on the Type I error rate, that is, the significance level α

For the pair of hypotheses: $H_0 : \mu = \mu_0$ versus $H_a : \mu > \mu_0$

It is intuitive that one should reject H_0, in support of the H_a, when the $\bar{X} > \mu_0$. Equivalently, this means to reject H_0 when the test statistic Z_0 is larger than certain positive value c ($Z_0 > c$).

The question is what is the exact value of c, which can be determined based on the significance level α, i.e., how much Type I error we would allow ourselves to commit.

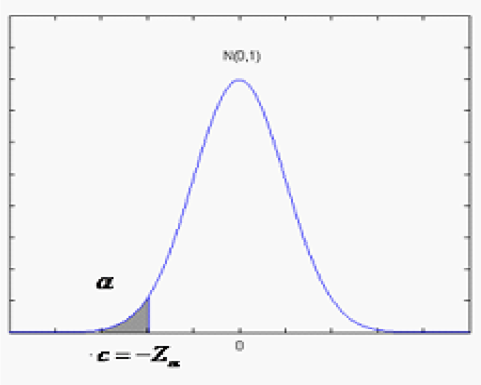
Hypothesis test for $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$



$\therefore$ At the significance level α, we will reject H_0 in favor of H_a if $Z_0 \geq Z_\alpha$

Hypothesis test for $H_0 : \mu = \mu_0$ versus $H_a : \mu < \mu_0$

Test statistic: $Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$



$\therefore$ At the significance level α, we will reject H_0 in favor of H_a if $Z_0 \leq -Z_\alpha$

Hypothesis test for $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$

Test statistic : $Z_0 = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

$P(|Z_0| \geq c \mid H_0) = \alpha$

$\Longleftrightarrow P(Z_0 \geq c \mid H_0) + P(Z_0 \leq -c \mid H_0) = \alpha$

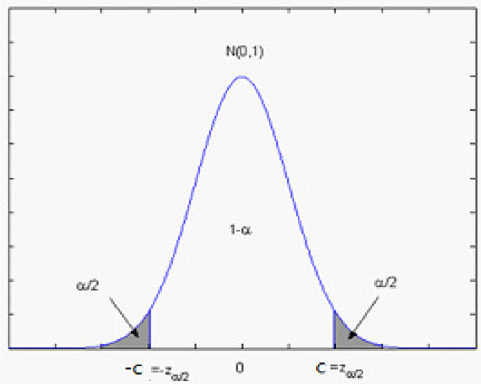
$2P(Z_0 \geq c \mid H_0) = \alpha$

$c = Z_{\frac{\alpha}{2}}$

$\therefore$ At the significance level α, we will reject H_0 in favor of H_a if

$|Z_0| \geq Z_{\frac{\alpha}{2}}$

Hypothesis test for $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$



P-value

- We have just discussed the **"rejection/critical region"** approach for decision making.
- There is another approach for decision making, it is the **"p-value"** approach.

Definition: p-value is the probability that of observing a test statistic value that is as extreme, or more extreme, than the one we observed.

- Reject H_0 in favor of H_a if p-value $\leq \alpha$.

Summary

<table>
<tr><td>$H_0 : \mu = \mu_0$
$H_a : \mu > \mu_0$</td><td>$H_0 : \mu = \mu_0$
$H_a : \mu < \mu_0$</td><td>$H_0 : \mu = \mu_0$
$H_a : \mu \neq \mu_0$</td></tr>
<tr><td colspan="3">Observed value of test statistic $Z_0 = \frac{\bar{X}-\mu_0}{\sigma / \sqrt{n}} \overset{H_0}{\sim} N(0,1)$</td></tr>
<tr><td colspan="3">Rejection region : we reject H_0 in favor of H_a at the significance level α if</td></tr>
<tr><td>$Z_0 \geq Z_\alpha$</td><td>$Z_0 \leq -Z_\alpha$</td><td>$|Z_0| \geq Z_{\alpha/2}$</td></tr>
<tr><td>p-value= $P(Z_0 \geq z_0 | H_0)$</td><td>p-value= $P(Z_0 \leq z_0 | H_0)$</td><td>p-value
$=P\left(|Z_0| \geq |z_0| \,|H_0\right)$
$= 2 * P\left(Z_0 \geq |z_0| \,|H_0\right)$</td></tr>
<tr><td>the area under $N(0,1)$ pdf to the right of z_0</td><td>the area under $N(0,1)$ pdf to the left of z_0</td><td>twice the area to the right of $|z_0|$</td></tr>
</table>

Read entire Chapter 7