Statistics ST2334 Short Recap: About a Mean

DONT COUNT ON THIS ALONE

2011/2012 Semester 2

Random sample

- Suppose we have a random sample $X_1, X_2, \ldots, X_n$. That is, they are independent and identically distributed with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$.
- We think of this as a simple random sample in a very large population.
- There are several objectives that we may be interested in; and they all rely on the sample mean:

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

Some Objectives

They include

1. Determine how the sample mean behaves given population parameters. E.g. what happens when you repeatedly make a bet?
2. Give a point estimate for μ given the sample, and error estimate. E.g. what is the average height of students in NUS? How sure are you?
3. Form a confidence interval for μ. E.g. what's a plausible range for μ based on the sample?
4. Determine sample size needed for a desired error level or interval width.
5. Hypothesis testing. E.g. Can we reject the current belief based on new evidence?

Sample Mean

- We first note that the sample mean

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

has

$$\begin{aligned} E(\bar{X}) &= E(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)) \\ &= \frac{1}{n}(E(X_1) + E(X_2) + \ldots + E(X_n)) \\ &= \mu, \text{ since } E(X_i) = \mu \end{aligned}$$

and

$$\begin{aligned} Var(\bar{X}) &= \frac{1}{n^2} Var(X_1 + X_2 + \ldots + X_n) \\ &= \frac{1}{n^2}(Var(X_1) + \ldots + Var(X_n)), \text{ by independence} \\ &= \frac{\sigma^2}{n}, \text{ since } Var(X_i) = \sigma^2 \end{aligned}$$

Normalized Sample Mean

- To give a more general description, we use the normalized version of the sample mean by subtracting it's mean and dividing by its standard deviation.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

  - This is just shifting and scaling $\bar{X}$ so that it has mean 0 and variance 1.
  - The shape of its distribution does not change.

Ok so what about the shape of the distribution?

- Now if our population is normal, that is each of our sample X_i follows the normal distribution, then $\bar{X}$ is also normal.
- If our population is unknown, but we have a large enough sample ($n \geq 30$), CLT tells us $\bar{X}$ is normal.
- In these cases, we therefore conclude that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Related forms

- Sometimes, σ is not available, and we use

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

  as a substitute.

- If n is large, this is a good estimate and we still have

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim N(0,1)$$

- If n is small but we know X is normal,

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

Summary

- We can summarize the above cases to

Case	σ	n	Population	Statistic	E
I	known	any	Normal	$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
II	known	large	any	$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
III	unknown	small	Normal	$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$	$t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$
IV	unknown	large	any	$Z = \frac{\bar{X}-\mu}{s/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$

Back to the objectives

1. We now know the distribution of the sample mean given population parameters. Can find probabilities etc.
2. We use $\bar{X}$ to estimate μ,
   - the standard error (SE) is $\sigma/\sqrt{n}$, the standard deviation of $\bar{X}$.
   - the maximum error E with probability 1-α is given by the table.
3. The (1-α) CI for μ is $\bar{X} \pm E$.
4. We rearrange the expression for E and solve for n.
5. We use the normalized sample mean as our test statistic.

Key Terms involving α

- Confidence Level $(1 - \alpha)$
  - This is for after estimates are made. e.g. we are 95% confident the population mean is in (a, b).
  - It is the probability of the *procedure* being correct, not the particular estimate.
- Significance Level α
  - Property of a hypothesis test.
  - It is the probability of making a Type I error when using said test.
  - It is an attempt to quantify how "significant" the result of a successful null hypothesis rejection.
  - The lower the significance level, the more confident that you correctly rejected the null.

Key Terms involving α

- Rejection Region
  - Based on the significance level α and the distribution of the test statistic.
  - The region depends on the alternative hypothesis. It is where the test statistic is deemed too extreme assuming the null (and more reasonable assuming the alternative).
  - The probability of the test statistic lying in the rejection region under the null hypothesis is α.
- *p*-value
  - The probability of observing your statistic or more "extreme" data, under the null hypothesis.
  - It is sometimes called the observed significance level.
  - The smaller the *p*-value, the more "unlikely" the null is. (Note that whether the null hypothesis is true is not random.)
  - The *p*-value is smaller than the significance level α if and only if the test statistic is in the rejection region.