Session 5: Inference using sample statistics

Susan Thomas

http://www.igidr.ac.in/~susant

susant@mayin.org

IGIDR

Bombay

Recap

- Samples and populations
- Parameters vs. statistics
- Sampling distribution of statistics
- Central Limit Theorem (CLT)
- Testing the CLT using the mean

Goals

- Applying the CLT: understanding the sample mean and it’s link with the population and the population mean.

Recap: Central Limit Theorem

Stating the CLT

- If the sample size is sufficiently large,
- the sample mean $\bar{x}$ has a sampling distribution that is approximately normal.

This is *irrespective* of the distribution of the underlying RV.

How large a sample?

Rule of thumb amongst statisticians: assuming the random variables are independant, $n = 30$ used to be taken as “sufficiently large”.

Properties of the sampling distribution of $\bar{x}$

Say $\bar{x}$ is the mean of a sample from a population, which has a mean population parameter μ and standard deviation parameter σ.
Then, we find that the sampling distribution of $\bar{x}$ has the following properties:

1. If $\mu_{\bar{x}}$ is the mean of the sampling distribution:

$$\mu_{\bar{x}} = \mu$$

2. If $\sigma_{\bar{x}}$ is the standard deviation of the sampling distribution:

$$\sigma_{\bar{x}} = \sigma/\sqrt{n}$$

Where n is the size of the sample.

Applying the CLT

Linking statistics and population parameters

The CLT implies that conditional on the sample being large enough, we can use the sample statistics to infer the population parameters.

Questions that we can answer

- Questions about the sample: Given that I know the population, how probable is the observed sample?
- Questions about the population: If we have a randomly selected sample, what can we say about the population?

Q1: Questions about the sample

- I calculate the returns calculated over a 45-day window before the day of budget announcement. I find that it is an average of 5% over the last 15 years. The σ of this 15-point sample is 7.5%.
- Returns in finance are approximated by a normal distribution. On average, the returns seen over 45-day windows in India in the past have been around 2.5%, with a volatility of 8%.
- **Question:** Are the budget period returns are the same as returns at other times of the year, or are they unusually high?
- **Statistical question:** How likely is it that my sample of budget returns could be drawn from a N(2.5,8)?

Q2: Questions about the population

- I know that the average height of a random sample of 20-30 year old men in Bombay city is 5.6 ft with a σ of 0.3 ft. This is calculated from a sample of 5500 men.
- We assume that this sample was randomly chosen.
- **Question:** What is the possible value of the mean height of 20-30 year old men in India?
- **Statistical question:** What is the mean and the standard deviation of this distribution?

Point estimates from sampling distributions

- The CLT tells us that under conditions of large sample sizes, $\bar{x}$ can be used as a "point estimate" of the population μ.
- However, any one sample gives a different "point estimate" of $\bar{x}$ – how reliable are these sample point estimates?
- The parameter estimate, therefore, ends up as a combination of a sample point estimate *and* a reliability measure.
- The reliability measure helps us with inference about the population parameter.

Inference for the statistic

- We know the sampling distribution of $\bar{x}$ is approximately normal, if the sample is large. (CLT) We also know that $\bar{\sigma}_{\bar{x}} = \sigma/\sqrt{n}$
- Then (for a normal distribution),

$$\bar{x} - 1.96\bar{\sigma}_{\bar{x}}, \bar{x} + 1.96\bar{\sigma}_{\bar{x}}$$

  captures approximately 95% of the values of sample means possible.
- 95 is called the **confidence level**, and the range of values is called the **confidence interval** for the statistic.

How it is used

- What we can say: μ lies with 95% probability between $\bar{x} - 1.96\bar{\sigma}_{\bar{x}}$, and $\bar{x} + 1.96\bar{\sigma}_{\bar{x}}$.
- NOTE: This does not mean that the 95% range of μ will be the same as the 95% range of $\bar{x}$ – just that μ is highly likely to fall somewhere within this range.

Testing concepts: population mean estimate

Suppose a random sample of 30 observation from the population of sale prices yielded the following statistics:

$$\bar{x} = 63,560, \sigma_x = 34,870$$

What is the 95% confidence interval for μ based on this sample?

Testing concepts: population mean estimate

Suppose a random sample of 30 observation from the population of sale prices yielded the following statistics:

$$\bar{x} = 63,560, \sigma_x = 34,870$$

What is the 95% confidence interval for μ based on this sample?
From CLT, σ_x is a good approximation for σ in large samples. Then,

$$63,560 \pm 1.96\frac{34,870}{\sqrt{30}} = 51,082 \;\; 76,038$$

gives us the range of values for μ with 95% confidence.

Testing concepts: population mean estimate

Housing sales in the city showed that the average ratio of sales price to appraised value in a sample of 50 sales showed the following statistics: $\bar{x} = 1.56, \sigma_x = 0.46$.
What is the mean ratio for all the houses sold in that year, using a 99% confi dence interval? (The 99% cutoff for the standard normal distribution is 2.58.)

Testing concepts: population mean estimate

Housing sales in the city showed that the average ratio of sales price to appraised value in a sample of 50 sales showed the following statistics: $\bar{x} = 1.56, \sigma_x = 0.46$.

What is the mean ratio for all the houses sold in that year, using a 99% confi dence interval? (The 99% cutoff for the standard normal distribution is 2.58.)

Then the estimate for the population will be somewhere between $1.56 \pm 2.58\frac{0.46}{\sqrt{50}} = 1.337, 1.675$.

With 99% confidence, we can be sure that the mean value of the ratio falls between 1.3 and 1.7. Interesting inference: in general, houses that year sold for prices higher than their appraised value – real estate is in a boom time here.

Testing concepts: population mean estimate

What if the sample statistics of $\bar{x} = 1.56, \sigma_x = 0.46$ had been collected from a sample of $n = 100$?

Testing concepts: population mean estimate

What if the sample statistics of $\bar{x} = 1.56, \sigma_x = 0.46$ had been collected from a sample of $n = 100$?

Then the estimate for the population will be somewhere between

$$1.56 \pm 2.58 \frac{0.46}{\sqrt{100}} = 1.44, 1.67$$

With a larger sample, the range of values for the mean ratio has dropped – we have become even more certain of the value of the population mean, even while the level of confidence of the test remains the same!

Questions about the sample, Q1

- Budget period $\bar{r} = 5\%$, $\sigma_r = 7.5\%$, $n = 15$
- Population $\mu_r = 2.5\%$, $\sigma = 8\%$

How likely is it to have observed the mean sample budget returns wrt population returns?

Questions about the sample, Q1

- Budget period $\bar{r} = 5\%$, $\sigma_r = 7.5\%$, $n = 15$
- Population $\mu_r = 2.5\%$, $\sigma = 8\%$

How likely is it to have observed the mean sample budget returns wrt population returns?

1. We first create what the sample mean's distribution ought to look like: $\mu_r = 2.5, \sigma = (8/\sqrt{15}) = 2.07$
2. The 95% interval for the standardised population is $2.5 \pm 1.96 * 2.07 = -1.56, 6.56$.
3. 5% falls in the 95% range of the population.

So we say that the budget returns is not atypical wrt rest of the population returns.

Questions about the sample, Q1

- Budget period $\bar{r} = 5\%$, $\sigma_r = 7.5\%$
- But what if we had only the sample, and did not have the population parameters?

Could we make a statement about how likely is it to have observed the mean budget returns are wrt population returns?

Questions about the sample, Q1

- Budget period $\bar{r} = 5\%$, $\sigma_r = 7.5\%$
- But what if we had only the sample, and did not have the population parameters?

Could we make a statement about how likely is it to have observed the mean budget returns are wrt population returns?

Since the sample is the only set of observations available, we might want to use the bootstrap – sampling with replacement – to recreate the sampling distribution of the sample.

Questions about the population, Q2

If a sample has the sampling statistics of $\bar{h} = 5.6$, $\sigma_h = 0.3$, $n = 5500$, what is the possible value of the mean height of 20-30 year old men in India?

Questions about the population, Q2

If a sample has the sampling statistics of $\bar{h} = 5.6$, $\sigma_h = 0.3$, $n = 5500$, what is the possible value of the mean height of 20-30 year old men in India?

1. The sampling distribution of the mean height is $\bar{h} = 5.6, \sigma_h = 0.3/\sqrt{5500}$.
2. The 95% range is $5.6 \pm 1.96 * 0.004 = 5.59, 5.61$

The mean height of the population of 20-30 year old men in India falls with 95% certainty between 5.59 and 5.61 ft.

Large vs. small samples

- All but one of these calculations were based on the assumption of large sample sizes.
- Two things go wrong with small samples:
  1. We can't assume that the sampling distribution is a normal. The sampling distribution depends upon the distribution of the underlying RV!
  2. We can't assume that the sample σ_x is a good approximation for σ.

Small sample solutions

- If the underlying RV distribution is “approximately normal”, then the confidence interval for μ is

$$\bar{x} \pm t_{\text{confi dence level}}\sigma_x$$

  So, $\bar{x} \pm t_{0.025}\sigma_x$, at a confidence level of 95%.
- The t-distribution is like the standard normal distribution, except that it has one more distributional parameter, the **degrees of freedom** (dof).
  dof = sample size (n) - 1
- At high dof, the t-distribution is the same as the std. normal distribution. At low dof, the t-distribution is flatter in the tails.

Correct solution for Q1

- The 95% range for the mean budget period returns is calculated using the cutoff using the t-distribution rather than the 1.96 from the normal distribution.
- $t_{0.025}(14)$
- $2.5 \pm 2.145 * 2.07 = -1.94, 6.94$
- This is a wider range of values – indicating a higher degree of uncertainty than if we had a “sufficiently large” sample.
- If the budget period returns had been 6.5%, we would have incorrectly infered that they were unusual using the normal distribution.

Applying CLT: Estimating $\mu_1 - \mu_2$

The problem

- There are two different populations.
- We want to estimate whether they have different means or not.
- We have to use a sample from each of these two populations.
- The samples are random, “large”, each with different sizes.

Example: Housing sales prices in Bombay vs. Delhi

- Random sample of 30 prices from Bombay:
  $\bar{m} = 52,356, \hat{\sigma}_m = 10,572$
- Random sample of 40 prices from Delhi:
  $\bar{d} = 66,491, \hat{\sigma}_d = 14,264$

These are sample values; We want the difference between the true mean sales prices for these two cities.

Difference between two means

- We assume that the Bombay prices are drawn from one population (with mean μ_b) and that Delhi prices come from another population (with mean μ_d).
- Our question is two-fold:
  1. What is the difference $\mu_b - \mu_d$?
  2. What is the 95% interval for $\mu_b - \mu_d$?
     If the 95% interval contains 0, it is likely that the average price of real estate in Bombay and Delhi are the same.
- We need the distribution of $\mu_b - \mu_d$ to answer these questions.

Distribution of $\mu_b - \mu d$

The distribution of $\mu_b - \mu d$ has

- $\mathrm{E}(\mu_b - \mu_d) = \mu_b - \mu_d$
- $\sigma = \sqrt{\frac{\sigma_b^2}{n_b} + \frac{\sigma_d^2}{n_d}}$

Then 95% confidence interval for $\mu_b - \mu_d$ will be

$$(\mu_b - \mu_d) \pm z_{0.05}\sqrt{\frac{\sigma_b^2}{n_b} + \frac{\sigma_d^2}{n_d}}$$

Applying CLT: from sample to population

- $\mathrm{E}(\mu_b) = \bar{x}_b$
- $\mathrm{E}(\mu_d) = \bar{x}_d$
- $\sigma_b \sim s_b$, sample standard deviation if the sample is "large" enough.
- $\sigma_d \sim s_d$
- The 95% confidence interval for $\mu_b - \mu_d$ will be

$$(x_b - x_d) \pm z_{0.05}\sqrt{\frac{s_b^2}{n_b} + \frac{s_d^2}{n_d}}$$

Solution to the sale price difference between Bombay and Delhi

- E($\mu_b - \mu_d$) = $(52,356 - 66,491) = -14,135$
- $\sqrt{\frac{10{,}572^2}{30} + \frac{14{,}264^2}{40}}$
- The 95% confidence interval is $-14,135 \pm 5,818.3$
- This is $-19,953.3, -8,316.7$.
- These numbers show that house sales in Bombay were at prices lower on average than those in Delhi.

With small samples

The changes in required assumptions are:

- The populations are relatively normal.
- The variances of the populations are **the same**!
- The samples are collected randomly and independantly from both the populations.

Then, the 95% confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{m} - \bar{d}) \pm t_{0.025} * \sqrt{\hat{\sigma}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

σ for the small sample

$\hat{\sigma}^2$ is a **pooled estimate of variance** and is calculated as:

$$\hat{\sigma}^2 = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{(n_1 + n_2 - 2)}$$

Testing concepts

Two samples have the following characteristics:

- $n_1 = 12, \bar{x}_1 = 10.6, \hat{\sigma}_1 = 2.4$
- $n_2 = 17, \bar{x}_2 = 9.5, \hat{\sigma}_2 = 4.7$

Then the difference between their true means are:

1. $\mu_1 - \mu_2 = (10.6 - 9.5) = 1.1$
2. $\hat{\sigma}^2 = \frac{(12-1)2.4^2+(17-1)4.7^2}{(12+17-2)} = 15.44$
3. $1.1 \pm 2.052\sqrt{15.44\left(\frac{1}{12} + \frac{1}{17}\right)} = 1.1 \pm 3.0 = -1.9, 4.1$

Applying CLT: Estimating population proportions

The problem

- There are two populations.
- We want to estimate what fraction of each population has a particular character:
  For example, what fraction of the house sales in either Bombay or Delhi was done by people below the age of 30.
- Further, we want to estimate how this fraction is different for Bombay as compared with Delhi.

Fraction of population, p

- Let's say that the fraction of house sales done by people below the age of 30 in the population of Bombay is p_b. This fraction for Delhi is p_d.
- We want to know $(p_b - p_d)$.
- We have a sample for each of the population in Bombay and Delhi, which gives us sample estimates $\hat{p}_b, \hat{p}_d$.
- In our example,
  $n_b = 200, n_d = 200, \hat{p}_b = 0.73, \hat{p}_d = 0.59$

Distribution of $p_1 - p_2$

- Assuming large samples, of size n_1, n_2,
- The mean of the difference is $p_1 - p_2$.
- $\sigma_{p_1 - p_2}$ is $\sqrt{\frac{\sigma_b^2}{n_b} + \frac{\sigma_d^2}{n_d}}$
- σ of the fraction is $p(1-p)$
- Then, $\sigma = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Example: Fractions in Bombay vs. Delhi

- $\mathrm{E}(p_1 - p_2) = \hat{p}_b - \hat{p}_d = 0.73 - 0.59 = 0.14$
- $s_1^2 = 0.73(1 - 0.73) = 0.1971$
- $s_2^2 = 0.59(1 - 0.59) = 0.2419$
- The 95% confidence interval for $(p_1 - p_2)$ is

$$0.14 \pm 1.96\sqrt{\frac{0.1971}{200} + \frac{0.2419}{200}}$$

- This is 0.05, 0.23

The fraction of below 30 year olds in Bombay making house purchases is higher than those in Delhi.

Problems to be solved

Q1: external audit fees

In the eighties, many US companies started their own audit departments to lower the costs of audits. A Harvard Business Review paper conducted a study of audit departments of 32 different companies. The author's main interest was to determine the effect of internal audit departments on external audit fees.

The mean external audit fee paid by the 32 companies in 1981 was USD 779,030 and the standard deviation was USD 1,083,162. What is highest fee that was paid in the 95% confidence range in that year?

Q2: Sugar content in food

Food manufacturers are required to list FDA estimates of the contents of the packaged product. Suppose you want to estimate the mean sugar content by weight in 16-oz boxes of corn flakes. The sample available to you is 100 boxes, from which you find that the sugar content has an average of 3.2 oz with a standard deviation of 0.5 oz.

1. What is the true mean sugar content in these 16-oz boxes of cornflakes (within a 90% confidence interval)?
2. How could the FDA reduce the width of the confidence interval? Are there any drawbacks to their doing this? What is the interpretation for the user of the information?

Q3: Small random sample

I have a sample of five measurements from a normally distributed population:

7 4 2 5 7

1. What is the 90% confidence interval for the mean?
2. What is the 99% confidence interval for the mean?
3. Do these two numbers make sense when we want to use them?

Q4: Branch offices

A company's branch office must make periodic shipments to branches in other states. In order to estimate the mean delivery time between two such offices, they selected five deliveries (at random) and recorded the time for each. The statistics were: $\bar{x} = 14.71$ hours and $\sigma_x = 0.87$ hours.

1. With 90% confidence, what is the true mean delivery time for all shipments between these offices?
2. What would the interval be if these were statistics from a sample of $n = 20$?
3. How can the offices uses these numbers to scale the information to other branches?

Q5: Mechanics of difference in means

Calculate the $\mu_{\bar{x}_1-\bar{x}_2}, \sigma_{\bar{x}_1-\bar{x}_2}$ for:

- $\bar{x}_1 = 150, \sigma_1^2 = 36, \bar{x}_2 = 140, \sigma_2^2 = 24, n_1 = n_2 = 35$
- $\bar{x}_1 = 125, \sigma_1^2 = 225, n_1 = 90, \bar{x}_2 = 112, \sigma_2^2 = 90, n_2 = 60$

Q6: Mechanics of difference in proportions

Calculate the $\mu_{\hat{p}_1-\hat{p}_2}, \sigma_{\hat{p}_1-\hat{p}_2}$ for:

- $\hat{p}_1 = 0.3, n_1 = 50, \hat{p}_2 = 0.4, n_2 = 30.$
- $\hat{p}_1 = 0.1, n_1 = 100, \hat{p}_2 = 0.05, n_2 = 100.$
- $\hat{p}_1 = 0.76, n_1 = 25, \hat{p}_2 = 0.96, n_2 = 25.$

Q7: Difference in proportions and sample sizes

Independant random samples produced the following results for two populations: $\hat{p}_1 = 0.44, \hat{p}_2 = 0.52$

- Find the 95% confidence interval for $(\hat{p}_1 - \hat{p}_2)$ if $n_1 = n_2 = 50$.
- Find the 95% confidence interval for $(\hat{p}_1 - \hat{p}_2)$ if $n_1 = n_2 = 500$.
- Find the 99% confidence interval for $(\hat{p}_1 - \hat{p}_2)$ if $n_1 = n_2 = 500$.

Q8: Quality of contractors

CERC wants to compare the safety records of two electrical contractors. They have inspected residences wired by each of the contractors and have recorded the number of residences that were electrically deficient and/or unsafe in the following table. What is the 95% confidence interval for the difference between the proportion of residences that are unsafe between the two contractors?

	Contractor A	Contractor B
Residences inspected	$n_A = 60$	$n_B = 70$
Number found unsafe	8	11

Q9: Small sample difference in means

Independent random samples from two normal populations showed the following:

Sample 1	Sample 2
5.3	6.8
5.1	7.4
6.9	7.9
7.4	9.1
9.7	

1. Find a 95% confidence interval for $(\mu_1 - \mu_2)$.
2. List any assumptions needed for the interval of part 1 to be valid.

Q10: More real estate differences between Bombay and Delhi

Recent random sample of purchase price for 100 new single family homes in Bombay and 80 in Delhi was done. Estimate the difference in the mean costs of such single-family homes in Bombay and Delhi using a 95% confidence interval using the results in the table (which shows costs in thousands of Rs.).

Bombay	Delhi
$\bar{x}_b = 94.8$	$\bar{x}_d = 73.4$
$s_b = 12.75$	$s_d = 16.8$

References

- Chapter 7, JAMES MCCLAVE and GEORGE BENSON. *Statistics for Business and Economics*. Dellen Publishing Company, 1991, 5th edition