BAN 602: Quantitative Fundamentals

Lecture 3: Central Limit Theorem and Confidence Interval

Agenda

1. Big Picture of Chapter 7: Sampling Distribution and Central Limit Theorem
2. Big Picture of Chapter 8: Confidence Interval
3. Example 1 (population mean - quantitative data)
4. Example 2 (population proportion - categorical data)

Big Picture of Chapter 7: Sampling Distribution & CLT

Sampling and Point Estimation	Sample Distribution	Properties of Point Estimators

• Population (finite and infinite) vs. Sample
• Sampling
  ○ Simple random sampling
  ○ Stratified random sampling
  ○ Cluster sampling
  ○ Systematic sampling
  ○ Convenience sampling
  ○ Judgment sampling
• Sample statistic (a sample characteristic: sample mean; sample proportion; sample standard deviation)
• Point estimation
	• Sampling Distribution (prob. Dist. of a sample statistic)
	  ○ **Sampling distribution of sample mean**
	  ○ **Sampling distribution of sample proportion**
	  ○ Sampling distribution of sample variance or sd
		• Unbiased
		• Efficiency
		• Consistency

Big Picture of Chapter 8: Interval Estimation

Basics of Interval Estimation	Population Mean	Population Proportion

• Interval estimate
• Margin of error
• Interval estimate of a population mean
• Interval estimate of a population proportion
• Interval estimate of difference of two population means/proportions (ch. 10)
  ○ What if more than 2 population means (ch. 13)
  ○ What if more than 2 population proportions (ch. 12)
• Interval estimate of a population variance & ratio of two population variances (ch. 11)
	• **Confidence level and confidence interval**
	• **When σ is known (generally speaking, use normal distribution)**
	• **When σ is unknown (generally speaking, use t distribution)**
	• Given desired margin of error, determine the appropriate sample size
		• **Interval estimation of population proportion**
		• **Determine the sample size**

Example 1 (Population Mean)

We want to study the annual cost of auto insurance.

- What would be the population of our interest? Let X be the random variable that describes the population. How can we interpret this random variable?

Example 1 (Population Mean)

We want to study the annual cost of auto insurance.

- What would be the population of our interest? Let X be the random variable that describes the population. How can we interpret this random variable?

  The population is the annual costs of all auto insurance policy.

  X represents the annual cost of a randomly selected auto insurance policy.

Example 1 (Population Mean)

We want to study the annual cost of auto insurance.

- What probability distribution does X follow? Note that this probability distribution describe the population.

Example 1 (Population Mean)

We want to study the annual cost of auto insurance.

- What probability distribution does X follow? Note that this probability distribution describe the population.

  We do not know the probability distribution of X. But we can make inference but the mean and variance/sd of this distribution.

Example 1 (Population Mean)

- How can we know about the (population) mean μ and the standard deviation σ?

Example 1 (Population Mean)

- How can we know about the (population) mean μ and the standard deviation σ?
    1. Randomly select n, called sample size, auto policies. Say, n is 25. Record their annual costs. Call this collection of 25 annual costs sample 1. Each of the 25 annual costs is called a sampling unit. Compute the mean and call it $\bar{X}_1$, a sample mean.
    2. Repeat the above step many times. We then have a collection of many sample means. These sample means are also random. Let $\bar{X}$ be the random variable representing the population of all sample means.

Example 1 (Population Mean)

- How can we know about the (population) mean μ and the standard deviation σ?
    3. Plot histogram with all the sample means collected and we can see the "empirical" probability distribution of $\bar{X}$. This distribution is called a sampling distribution (of sample mean). The sampling distribution also has its mean, $\mu_{\bar{X}}$, and standard deviation, $\sigma_{\bar{X}}$. The properties of the sampling distribution of sample mean are summarized as central limit theorem (CLT).

Example 1 (Population Mean)

- How can we know about the (population) mean μ and the standard deviation σ?
  4. Central limit theorem.
  - $\bar{X}$ approximately follows a normal distribution, i.e., the sampling distribution of sample mean is approximately normal.
  - The larger the sample size n is, the closer the sampling distribution is to a normal distribution.
  - $\mu_{\bar{X}} = \mu$, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

Example 1 (Population Mean)

- Now, let's assume the (population) mean $\mu = \$939$ and the (population) standard deviation $\sigma = \$245$, and $n = 25$. Can you describe the sampling distribution of $\bar{X}$?

Example 1 (Population Mean)

- Now, let's assume the (population) mean μ = $939 and the (population) standard deviation σ = $245, and n = 25. Can you describe the sampling distribution of $\bar{X}$?

The sampling distribution of sample mean is approximately normal, with $\mu_{\bar{X}}$, the mean of the sampling distribution being the same as the population mean, $939, and $\sigma_{\bar{X}}$, the standard deviation of the sample distribution (aka standard error) being 245/5 = $49.

Example 1 (Population Mean)

- What is the probability that a simple random sample of automobile insurance policies will have a sample mean within $25 of the population mean?

Example 1 (Population Mean)

- What is the probability that a simple random sample of automobile insurance policies will have a sample mean within $25 of the population mean?

Pr($\mu - 25 \leq \bar{X} \leq \mu + 25$) = pnorm(939+25, 939, 49) - pnorm(939-25, 939, 49) = 39%

Example 1 (Population Mean)

- In reality, we rarely know μ and σ. We instead apply the process reversely: use the sample mean to infer the population mean. And this process is called **statistical inference**. Now suppose we collect information on 25 auto insurance polices and their average annual cost is \$1000. For the time being, assume we still know σ = \$245 (We will handle the situation where σ is unknown later). What is the probability that the population mean is within \$50 of the sample mean?

Example 1 (Population Mean)

- To answer the previous question, let's consider the general case. We randomly sampled n auto policies and the mean price is $\bar{X}$. What is the probability that the population mean is within \$m (this is called margin of error) of the sample mean? This range of $\bar{X}$ $\pm$m is called a confidence interval and the probability $Pr(\bar{X} - m \leq \mu \leq \bar{X} + m)$ is called the confidence level.

Example 1 (Population Mean)

- To answer the previous question, let's consider the general case. We randomly sampled n auto policies and the mean price is $\bar{X}$. What is the probability that the population mean is within \$m (this is called margin of error) of the sample mean? This range of $\bar{X}$ $\pm$m is called a confidence interval and the probability $Pr(\bar{X} - m \leq \mu \leq \bar{X} + m)$ is called the confidence level.

$$\Pr(\bar{X} - m \leq \mu \leq \bar{X} + m)$$
$$= \Pr(\mu - m \leq \bar{X} \leq \mu + m)$$
$$= P(-m \leq \bar{X} - \mu \leq m)$$
$$= P\left(\frac{-m}{\sigma_{\bar{X}}} \leq \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \leq \frac{m}{\sigma_{\bar{X}}}\right)$$
$$= P\left(\frac{-m}{\sigma/\sqrt{n}} \leq z \leq \frac{m}{\sigma/\sqrt{n}}\right)$$

Example 1 (Population Mean)

- In reality, we rarely know μ and σ. We instead apply the process reversely: use the sample mean to infer the population mean. And this process is called **statistical inference**. Now suppose we collect information on 25 auto insurance polices and their average annual cost is $1000. For the time being, assume we still know σ = $245 (We will handle the situation where σ is unknown later). What is the probability that the population mean is within $50 of the sample mean?

$$\Pr(1000 - 50 \leq \mu \leq 1000 + 50)$$

$$= P\left(\frac{-50}{245/\sqrt{25}} \leq z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{50}{245/\sqrt{25}}\right)$$

$$= \text{pnorm}(50/49) - \text{pnorm}(-50/49) = 2^{*}\text{pnorm}(50/49) - 1 = 69.25\%$$

Example 1 (Population Mean)

- In practice, we typically have desirable confidence levels, with 90%, 95%, and 99% being the most commonly used ones. We, instead, want to find the corresponding margin of error and the resulting confidence interval.
- Once again, assume σ, the population standard deviation is known. We randomly sampled n auto policies and the mean price is $\bar{X}$.
- Suppose the confidence level we want is 1-α. (α is called significance level, which we will use extensively later on. If the confidence level is, say 90%, then the significance level is 10%, vice versa.) What would be the margin of error that provides the confidence level of 1-α? And what would be the confidence interval that provides the confidence level of 1-α?.

Example 1 (Population Mean)

- Suppose the confidence level we want is 1-α. What would be the margin of error (m) that provides the confidence level of 1-α? And what would be the confidence interval that provides the confidence level of 1-α?.

Pr($\bar{X} - m \le \mu \le \bar{X} + m$) = 1-$\alpha$ (previously we know m and try to find 1-α)

$P\left(\frac{-m}{\sigma/\sqrt{n}} \le \mathrm{z} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{m}{\sigma/\sqrt{n}}\right)$ = 1-α

$P\left(\mathrm{z} \le \frac{m}{\sigma/\sqrt{n}}\right)$ = 1-$\alpha/2$ (z is a standard normal random variable)

$\frac{m}{\sigma/\sqrt{n}} = F^{-1}_{standard\ normal}(1-\frac{\alpha}{2})$=qnorm(1-$\alpha/2$)

$m = F^{-1}_{s.n.}\left(1-\frac{\alpha}{2}\right) * \sigma/\sqrt{n}$= qnorm(1-$\alpha/2$)* $\sigma/\sqrt{n}$

1-α Confidence interval: [$\bar{X} - m$, $\bar{X} + m$]

Example 1 (Population Mean)

- Now suppose we collect information on 25 auto insurance polices and their average annual cost is $1000; σ = $245. What would be the margin of error that provides a 95% confidence level? And what would be a 95% confidence interval?. How can we interpret this confidence interval?

Example 1 (Population Mean)

- Now suppose we collect information on 25 auto insurance polices and their average annual cost is $1000; σ = $245. What would be the margin of error that provides a 95% confidence level? And what would be a 95% confidence interval? How can we interpret this confidence interval?

m = qnorm(1-5%/2, 0, 1)*245/5 = 1.96*49 = 96.04

95% confidence interval: [1000-96.04, 1000+96.04] = [904, 1096]

There is 95% chance (We are 95% confident) that the population mean will be within the range of 904 and 1096.

Example 1 (Population Mean)

- Sometimes we have desirable margin of error and confidence level. We want to find the corresponding sample size that can help us achieve the desirable m and 1-α. How?

Example 1 (Population Mean)

- Sometimes we have desirable margin of error and confidence level. We want to find the corresponding sample size that can help us achieve the desirable m and 1-α. How?

Essentially, in $m = F_{s.n.}^{-1}\left(1-\frac{\alpha}{2}\right) * \sigma/\sqrt{n}$, we always know m, σ, and α. We want n.

$$n = \left(F_{s.n.}^{-1}\left(1-\frac{\alpha}{2}\right) * \sigma/m\right)^2 = \left(qnorm\left(1-\frac{\alpha}{2}\right) * \sigma/m\right)^2$$

Example 1 (Population Mean)

- Now suppose the population standard deviation of annual cost of auto insurance polices σ = \$245. What sample size can ensure that the margin of error of 99% confidence level is \$50?

$$n = \left(qnorm\left(1 - \frac{\alpha}{2}\right) * \sigma/m\right)^2 = \left(qnorm\left(1 - \frac{1\%}{2}\right) * 245/50\right)^2$$

$$= (2.576 * 4.9)^2 = 160\ (159.3)$$

Example 1 (Population Mean)

- What are the impacts of large sample size n?

$$m = F_{s.n.}^{-1}\left(1 - \frac{\alpha}{2}\right) * \sigma/\sqrt{n}$$

Example 1 (Population Mean)

- Now what if σ is unknow, which typically is the case?

Example 1 (Population Mean)

- Now what if σ is unknow, which typically is the case?

  statistic $z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ no longer available

  $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$

  $m = F_t^{-1}\left(1-\frac{\alpha}{2}, n-1\right) * s/\sqrt{n}$= pt(1-$\alpha/2$)* $s/\sqrt{n}$

Example 2 (Population Proportion)

The president of Doerman Distributors, Inc., wants to know the percentage of the firm’s orders that come from first-time customers. To that end, a random sample of 100 orders is collected among which 30% of the firm’s orders come from first-time customers.

- What is the population parameter of interest? What do we know about this population parameter?

Example 2 (Population Proportion)

- What is the population parameter of interest?

  p: proportion of first-time customers' orders among all the firm's orders

- What do we know about this population parameter?

  consider random variable $Y = 1$ if it's an order from first-time customer and $Y = 0$, otherwise.

  $E(Y) = p$; $Var(Y) = p(1-p)$.

Example 2 (Population Proportion)

- We use sample statistic, sample proportion $\bar{p}$, to study p. What do we know about $\bar{p}$?

Example 2 (Population Proportion)

- We use sample statistic, sample proportion $\bar{p}$, to study p. What do we know about $\bar{p}$?

$\bar{p}$ is approximately normally distributed with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$. Alternatively, $\bar{p} \sim norm(p, \sqrt{\frac{p(1-p)}{n}})$ or $z = \frac{\bar{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim norm(0,1)$.

Example 2 (Population Proportion)

The president of Doerman Distributors, Inc., wants to know the percentage of the firm’s orders that come from first-time customers. To that end, a random sample of 100 orders is collected among which 30% of the firm’s orders come from first-time customers.

- If the confidence level is 90%, what is the margin of error and what is a confidence interval for the population proportion?
- How can we interpret this confidence interval?

Example 2 (Population Proportion)

- If the confidence level is 90%, what is the margin of error and what is a confidence interval for the population proportion?

$$m = F_{s.n.}^{-1}\left(1 - \frac{\alpha}{2}\right) * \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = qnorm(0.95) * \sqrt{\frac{0.3*0.7}{100}} = 1.645 * 4.58\% = 7.54\%.$$

Confidence interval: [30%-7.54%, 30%+7.54%] = [22.46%, 37.54%]

- How can we interpret this confidence interval?

There is 90% chance that the population proportion is b/w 22.46% and 37.54%.

Example 2 (Population Proportion)

You are probably not happy with so wide a confidence interval. Suppose you want a confidence level of 95% (1-α) and a margin of error 3% (m). What sample size will be needed?

Example 2 (Population Proportion)

You are probably not happy with so wide a confidence interval. Suppose you want a confidence level of 95% (1-α) and a margin of error 3% (m). What sample size will be needed?

$$m = F_{s.n.}^{-1}(1-\alpha/2) * \sqrt{\bar{p}(1-\bar{p})/n} = qnorm(0.975) * \sqrt{0.3 * 0.7/100} = 1.645 * 4.58\% = 7.54\%.$$

$$n = \frac{\left(F_{s.n.}^{-1}(1-\alpha/2)\right)^2}{m^2} * \bar{p}(1-\bar{p}) \leq \frac{\left(F_{s.n.}^{-1}(1-\alpha/2)\right)^2}{4m^2}$$

$$n \leq \frac{\left(qnorm(0.975)\right)^2}{4*(0.03)^2} = 1067 \; or \; 1068 \; .$$

Sample Size Requirements (Population Proportion)

	m = 1%	m = 3%	m = 5%
90%	6764	752	271
95%	9604	1068	385
99%	16588	1844	664

Quiz 3 Part 1

1. Write pmf of a discrete random variable.
2. Write cdf of a discrete random variable.
3. Compute expectation, covariance, etc. (formula sheet for exam 1).
4. Independence
5. Relationship between correlation and independence.

Quiz 3 Part 2

1. Describe the sampling distribution of sample mean or proportion.
2. Compute the point estimate and the standard error.
3. Write down R code for computing margin of error.
4. Provide confidence interval and its interpretation.
5. Write down R code that computes the sample size needed for desirable confidence level and margin of error.