Second Course in Statistics: lecture 12

Review: Sampling distributions

- Sampling distribution of sample mean, proportion and variance

Statistical inference

- Purpose of statistical inference
- Finding a good estimator
- Confidence interval estimation
- Hypothesis testing

Statistical inference

Purpose of statistical inference

- At the heart of making sense of data, we are usually interested in making an inference about a population based on information contained in the sample.
- Populations are unknown, but are characterized by numerical measurements called parameters and sometimes are associated with certain probability distributions. A population follows a certain probability distribution with unknown parameters, typically mean value and standard deviation or population proportion. Discovery of these parameters will disclose the nature of the population.
- The purpose of statistical inference is to infer (predict, or estimate) the parameters through the sample data collected from the population.

Statistical inference

Types of statistical inference

There are two most prominent types of formal statistical inference.

1. Confidence interval estimation of a population parameter
2. Hypothesis testing: assessing a claim about the population parameter

Both types of inference need
good estimator and sampling distribution of the estimator
Both types of inference answer the question:
What would happen if we used the inference method many times.

Statistical inference

Finding a good estimator

Definition

When a statistic is used to estimate the population parameter, it is called an **estimator**.

Two basic criteria to evaluate estimators:

1. A good estimator is unbiased. (unbiasedness means that the expected value of an estimator is equal to the population parameter which this estimator is going to predict.)
2. A good estimator has smaller variance. The standard deviation of an estimator is usually called standard error.

Statistical inference

Finding a good estimator

Example

Mean is one of the measurement for central tendency in the population. There are other ways to measure the central tendency, such as median and mode. One natural choice is to use sample mean to estimate population mean. Why not choose the others?

Answer

Sample mean

1. is an unbiased estimator
2. has smaller variability than the other two.

Statistical inference

Finding a good estimator

Example

Sample variance defined as $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is chosen as a statistic to estimate the population variance and why not choose

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Answer

- $ES^2 = \sigma^2$ S^2 is unbiased.
- $ES_1^2 = \frac{n-1}{n}\sigma^2$ S_1^2 underestimates the population variance.

Statistical inference

Confidence interval estimation for population mean

Aim

Answer the question of where true population mean value is located

Derivation

Sampling distribution of sample mean.
According to Central Limit Theorem (CLT)

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

where μ is the parameter that needs to be estimated under the assumption that population σ^2 is known.

Statistical inference

Confidence interval estimation for population mean

Definition

Confidence level is defined as

$$c = 1 - \alpha$$

and confidence interval for population μ when σ is known is

$$\left(\bar{X} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \quad \bar{X} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right)$$

In particular

- 90% confidence interval is $\bar{X} \pm 1.645\sigma/\sqrt{n}$
- 95% confidence interval is $\bar{X} \pm 1.96\sigma/\sqrt{n}$
- 99% confidence interval is $\bar{X} \pm 2.576\sigma/\sqrt{n}$

Statistical inference

Confidence interval estimation for population mean

Example

The SAT tests are widely used as measures of readiness for college study. There are two parts, one for verbal reasoning ability (SATV) and one for mathematical reasoning ability (SATM). In 2003, 1,406, 324 college-bound senior took the SAT. Their mean SATV score was 507 with a standard deviation of 111. For SATM the mean is 519 with standard deviation of 115.

You want to estimate the mean SATM score for more than 385,000 high school seniors in California. At considerable effort and expense, you give the test to a random sample of 500. The mean score for your sample is 461. Explain the mean score in the population of 385,000 high school students.

Statistical inference

Confidence interval estimation for population mean

Solution

- Known population: SATV and SATM scores of 1,406, 324 high school seniors

$$SATV \sim (507,\ 111^2) \qquad SATM \sim (519, 115^2)$$

- Unkown population: SATM scores of 385,000 high school seniors in California.
- 95% confidence interval for mean score of 385,000 high school seniors in California is

$$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}} \Longrightarrow 461 \pm 1.96\frac{115}{\sqrt{500}}$$
$$\Longrightarrow (450.92,\ 471.08)$$

Statistical inference

Confidence interval estimation for population mean

Interpretation

- Among $100(1-\alpha)\%$ repeated sampling, the constructed confidence interval covers th true population mean. In other words, there is a $100\alpha\%$ chance that the constructed interval does not include true population mean.
- For SAT example, we can be 95% sure that mean score of California high school seniors is located within the interval $(450.92, 471.08)$

Statistical inference

Confidence interval estimation for population mean

Remark

- For a sample of size n, the confidence interval expands as the confidence level increases
- For a certain confidence level, confidence interval shrinks as the sample size increases.
- Confidence interval estimation is a rather general concept applying to all parameter estimation, not restricting to mean.
- Confidence interval is not necessarily symmetrical around sample value, such as interval estimation for population variance.

Statistical inference

Confidence interval estimation for population mean

Example

An archaeologist discovers a short manuscript in an ancient language which he recognises but cannot read. There are 30 words in the manuscript and they contain a total of 198 letters. There are two written versions of the language. In the early form of the language the mean word length is 6.2 letters with standard deviation 2.5; in the late form certain words were given prefixes, raising the mean length to 7.6 letters but leaving the standard deviation unaltered. The archaeologist hopes the manuscript will help him to date the site.
Construct a 95% confidence interval for the mean word length of the language from where your sample is drawn. Would you recommend to the archaeologist that this manuscript belong to early form of the language?

Statistical inference

Confidence interval estimation for population mean

Solution

- Known population: word length of early form language defined as X and word length of late form language defined as Y, then

$$X \sim (6.2,\ 2.5^2) \qquad Y \sim (7.6,\ 2.5^2)$$

- Sample is taken from one of these populations but we don't know which one, so confidence interval estimation discloses which population the sample is drawn from.
- Sample mean is $198/30 = 6.6$
- 95% confidence interval for population mean is

$$6.6 \pm 1.96 \times 2.5/\sqrt{30} \implies (6.144,\ 7.056)$$

Statistical inference

Confidence interval estimation for population mean

Solution continued

- The mean word length of early form language 6.2 is in the estimated 95% confidence interval (6.144, 7.056) but 7.6 which is mean word length of late form is not in the interval. We conclude that we are 95% certain the sample was taken from early form language.

Statistical inference

Hypothesis testing

Aim

Answer the question: ” Is the parameter equal to (or larger than or less than) a specified value?”
It is a formal procedure for comparing observed data with a hypothesis we want to assess.

Test procedure

1. State the hypothesis: null hypothesis denoted as H_0 and alternative hypothesis denoted as H_a (or H_1). Alternative hypothesis is also called research hypothesis.
2. Specify sampling distribution of the test statistic and calculate observed test statistic (denoted as T.S.) according to the random sample.
3. Specify rejection region (denoted by R.R.) according to level of significance and H_a
4. Compare test statistics and rejection region and draw a proper conclusion.

Statistical inference

Hypothesis testing

Null hypothesis

- The statement being tested in a statistical test is called the null hypothesis. The hypothesis testing is designed to assess the strength of the evidence against the null hypothesis. Usually null hypothesis is a statement of ”no effect”, ”no difference”, or ”no change” and the unknown parameter is assumed to be a certain value.

Alternative hypothesis

- Alternative hypothesis should express the hopes or suspicions we bring to the data which might contradict H_0. Care must be taken when we state the alternative hypothesis. The contrary of ”no difference” might be smaller or larger, or both. The specification of alternative hypothesis affects rejection region.

Statistical inference

Hypothesis testing

Test Statistics

The test is based on a statistic that estimates the parameter specified in the hypothesis. Usually this is the same estimator we would use in the confidence interval for the parameter. A test statitic is constructed under the H_0.

Statistical inference

Hypothesis testing

Test Statistics for population mean

1. Sample mean $\bar{X}$ as an estimator. By CLT

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$$

and we usually sketch the pdf curve for $\bar{X}$ and identify the value of $\bar{X}$ from the sample in the graph.

2. Standardized statistic

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

as an estimator and sketch a standard normal curve and identify z value calculated from the sample in the graph

Statistical inference

Hypothesis testing

Rejection region

- When H_0 is true, we expect that the realized value of the estimator found in test statistic to be near the parameter value specified in H_0.
- Values of the estimator far from the parameter value specified by H_0 give the evidence against H_0.
- The cutoff point for "near and far" is specified by level of significance. The cut-off point is called critical value.
- Rejection region is composed of the values of the estimator starting from critical value and further away from assumed parameter.
- Alternative hypothesis determines the direction of rejection region. Rejection region can be left side or right side or both sides.

Level of significance

Definition

Level of significance is a cut-off criterion to specify how far the evidence is away from the parameter value (assumed under H_0) to be considered as evidence against H_0

Level of significance is a probability value and specifies the probability of the values of estimator which are unlikely under H_0.

Typical levels of significance are 5%, 1%, 10%. The level of significance can specify the unlikely events on left side, or right side or both.

Statistical inference

Hypothesis testing

Comparison and conclusion

- If the test statistic calculated from the sample data is in the rejection region, we conclude that we can reject H_0 at a certain level of significance.
- If the test statistic calculated from the sample data is not in the rejection region, we conclude that we cannot rejection H_0 at the level of significance.
- Be aware that we do not accept H_0 or H_a without further assessment.

Manuscript example continued

An archaeologist discovers a short manuscript containing 30 words and 198 letters in total. There are two written versions of the language. In the early form of the language the mean word length is 6.2 letters with standard deviation 2.5; in the late form certain words were given prefixes, raising the mean length to 7.6 letters but leaving the standard deviation unaltered. Assume that this manuscript was taken from early form of the language. Assess this claim by hypothesis testing at 5% level of significance.

Statistical inference

Hypothesis testing

Solution

1. Specify Hypotheses:
   - H_0: $\mu = 6.2$
   - H_a: $\mu > 6.2$
2. Test statistic :
   - $\bar{X} \sim N\left(6.2, \frac{2.5^2}{\sqrt{30}}\right)$ or
   - $\frac{\bar{X} - 6.2}{2.5/\sqrt{30}} \sim N(0, 1)$
   - Observed $\bar{x} = 198/30 = 6.6$ or
   - the observed $z = \frac{6.6 - 6.2}{2.5/\sqrt{30}} = 0.8764$

Statistical inference

Hypothesis testing

Solution continued

3 Rejection region:

- Under the sampling distribution of $\bar{X}$, rejection region is $X > 6.2 + 1.645 \times \frac{2.5}{\sqrt{30}} = 6.951$
- Under the sampling distribution of Z, rejection region is just $Z > 1.645$

4 Comparison and conclusion :

- Since the observed $\bar{x} = 6.2 < 6.951$, we cannot reject H_0 at 5% level of significance and we conclude that we haven't find critical evidence to prove that this manuscript is from late language form.
- The same conclusion is drawn when comparing the value of z.