

Overview of Statistical Concepts

Lecture 3

SIE 512 Spatial Analysis

Fall 2009

- Parameter Estimation
- Confidence Intervals
- Hypothesis Testing

Populations and samples

Population

Any collection of people, animals, plants or things about which we may collect data. It is the entire group we are interested in.

We make generalisations about a population from **samples**, that are meant to be representative of the population.

We use sample statistics to give information about a corresponding population parameter. The sample mean for a set of data gives information about the overall population mean.

Sample statistics and population parameters

We have a random variable we assume to be distributed according to some distribution $X \sim f(\theta)$

We want to use a sample of data to estimate the population parameter θ

Parameter: a value, usually unknown (so needs to be estimated), used to represent a certain population characteristic

Statistic: a quantity that is calculated from a sample of data

Sampling Distribution

Describes probabilities associated with a statistic for a random sample drawn from a population.

The sampling distribution is the probability distribution or probability density function of the statistic.

Derivation of the sampling distribution is the first step in calculating a confidence interval or carrying out a hypothesis test for a parameter.

Estimates and Estimators

Estimate: an indication of the value of an unknown quantity based on observed data.

Estimator: a quantity calculated from the sample data which is used to give information about an unknown quantity in the population - distinguished from the true value by using the 'hat' symbol

The sample mean is an **estimator** of the population mean.

For n independent normally distributed variables $X \sim N(\mu, \sigma^2)$

The sample mean has a sampling distribution $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

Sample mean is normally distributed with reduced variance

Central limit theorem

The central limit theorem states that given a distribution with a mean μ and variance σ^2 , the sampling distribution of the mean approaches a normal distribution with a mean (μ) and a variance σ^2/N as the sample size N , increases.

The interesting thing about the central limit theorem is that it does not matter what the original distribution is, the sampling distribution of the mean approaches a normal distribution.

Sampling Distributions

The sampling distribution of a sample statistic depends on:

- The underlying probability distribution of the random variable – determined by its population parameters
- The particular sample statistic – usually the mean or variance
- The sample size n – the larger the sample size, the smaller the spread of the sample distribution

Sampling errors

To address the uncertainty of an estimate of a parameter θ , an interval estimate is preferred over a point estimate.

One estimate of sampling uncertainty is described by the standard error – the standard deviation of the sample statistic

If t is the sample statistic and s_t is the sample standard deviation, the error interval is $t \pm s_t$

Specifies an interval likely to cover the population parameter

Confidence intervals

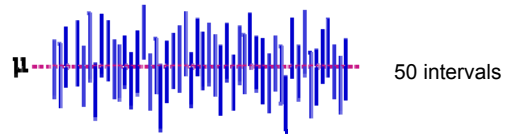
The most prevalent form of interval estimation is the **confidence interval** (a frequentist method) and credible intervals (a Bayesian method).

A **confidence interval**: an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data.

The **confidence level** is the probability that the interval produced by the method employed includes the true value of the parameter.

Confidence intervals and levels

When independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, a certain percentage (confidence level) of the intervals will include the unknown population parameter. This percentage can be 90%, 95%, 99%, 99.9%.



Calculating a Confidence Interval

The endpoints of the confidence interval are calculated from the sample, so they are statistics, functions of the sample and random variables themselves

CI for the sample mean

The sampling distribution for the sample mean tends in the limit of large N to

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

Test statistic is Z

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \quad \text{Normal distribution transformed to standard normal distribution} - N(0,1)$$

The (1- α) 100% CI for μ can be written as

$$\bar{x} - Z_c \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_c \frac{\sigma}{\sqrt{n}}$$

Calculating a Confidence Interval

Given the random variable Z with a standard normal distribution independent of the parameter μ to be estimated, it is possible to find numbers $-z$ and z , independent of μ , where Z lies in between with probability $1 - \alpha$, a measure of how confident we want to be

We take $1 - \alpha = 0.95$. So we have:

$$P(-z \leq Z \leq z) = 1 - \alpha = 0.95$$

The number z follows from:

$$\Phi(z) = P(Z \leq z) = 1 - \frac{\alpha}{2} = 0.975$$

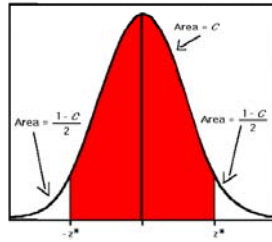
$$z = \Phi^{-1}(\Phi(z)) = \Phi^{-1}(0.975) = 1.96$$

Confidence Interval for sample mean

For a normal distribution, the confidence intervals correspond to percentages of the area of the normal density curve. A 95% confidence interval covers 95% of the normal curve -- the probability of observing a value outside of this area is less than 0.05.

Because the normal curve is symmetric, half of the area is in the left tail of the curve, and the other half is in the right tail.

For a confidence interval with level C , the area in each tail of the curve is equal to $(1-C)/2$. For a 95% confidence interval, the area in each tail is equal to $0.05/2 = 0.025$.



Confidence Interval

The value z^* representing the point on the standard normal density curve such that the probability of observing a value greater than z^* is equal to p is known as the upper p critical value of the standard normal distribution. If $p = 0.025$, the value z^* such that $P(Z > z^*) = 0.025$, or $P(Z \leq z^*) = 0.975$, is equal to 1.96.

For a confidence interval with level C , the value p is equal to $(1-C)/2$. A 95% confidence interval for the standard normal distribution, then, is the interval $(-1.96, 1.96)$, since 95% of the area under the curve falls within this interval.

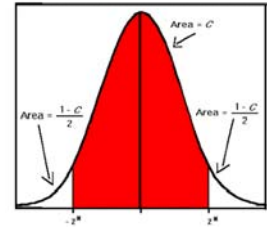


Table of Critical Values

α	$1 - \alpha$	Z_c	Description
0.50	0.50	0.68	50% C.I. \pm one "probable error"
0.32	0.68	1.00	68% C.I. \pm one "standard error"
0.10	0.90	1.65	90% C.I.
0.05	0.95	1.96	95% C.I. about \pm two standard errors
0.001	0.999	3.29	99.9% C.I. about \pm three standard errors

Table 5.1: Critical values for various common confidence levels

Example

The boiling temperature of a certain liquid is observed (in degrees Celsius) as follows: 102.5, 101.7, 103.1, 100.9, 100.5, and 102.2 for 6 different samples of the liquid. The sample mean is calculated to be 101.82. The standard deviation for this procedure is 1.2 degrees.

If the measurements follow a normal distribution, the sample mean will have the distribution $N(\mu, \frac{\sigma}{\sqrt{n}})$.

Since the sample size is 6, the standard deviation of the sample mean is equal to $1.2/\sqrt{6} = 0.49$.

For a population with unknown mean and known standard deviation, a confidence interval for the population mean, based on a simple random sample (SRS) of size n , is $\bar{X} \pm z^* \frac{\sigma}{\sqrt{n}}$ where z^* is the upper $(1-C)/2$ critical value for the standard normal distribution.

Example

The sample mean of the boiling temperatures was 101.82, with standard deviation 0.49. The critical value for a 95% confidence interval is 1.96. A 95% confidence interval for the unknown mean is $((101.82 - (1.96 \cdot 0.49)), (101.82 + (1.96 \cdot 0.49))) = (101.82 - 0.96, 101.82 + 0.96) = (100.86, 102.78)$.

As the level of confidence decreases, the size of the corresponding interval will decrease. For a 90% confidence level, $C = 0.90$, and $(1-C)/2 = 0.05$. The critical value z^* for this level is equal to 1.645, so the 90% confidence interval is $((101.82 - (1.645 \cdot 0.49)), (101.82 + (1.645 \cdot 0.49))) = (101.82 - 0.81, 101.82 + 0.81) = (101.01, 102.63)$

Confidence Intervals for Unknown Mean and Unknown Standard Deviation

Most often, the standard deviation for the population of interest is not known. In this case, the standard deviation is replaced by the estimated standard deviation s , - the **standard error**. Since the standard error is an estimate for the true value of the standard deviation, the distribution of the sample mean is no longer normal with mean μ and standard deviation σ/\sqrt{n} . Instead, the sample mean follows the **t distribution** with mean μ and standard deviation s/\sqrt{n} .

The *t* distribution is described by its **degrees of freedom**. For a sample of size n , the *t* distribution will have $n-1$ degrees of freedom. As the sample size n increases, the *t* distribution becomes closer to the normal distribution, since the standard error approaches the true standard deviation for large n .

Student's *t*-distribution

Confidence intervals and hypothesis tests use Student's *t*-distribution to cope with uncertainty resulting from estimating the standard deviation from a sample.

It is the basis of the popular Student's *t*-tests for the statistical significance of the difference between two sample means, and for confidence intervals for the difference between two population means.

Student's *t*-distribution

Student's *t*-distribution has the probability density function

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)},$$

where ν is the number of *degrees of freedom* and Γ is the Gamma function.

The distribution depends on ν , but not μ or σ ;

Differs from *Z* in that the exact standard deviation σ is replaced by the random variable

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}},$$

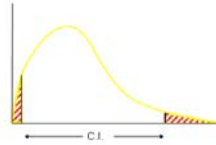
Confidence Intervals for Unknown Mean and Unknown Standard Deviation

For a population with unknown mean μ and unknown standard deviation, a confidence interval for the population mean, based on a simple random sample of size n , is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the upper $(1-C)/2$ critical value for the t distribution with $n-1$ degrees of freedom, $t(n-1)$.

Confidence Intervals for sample variance



The sample variance is the sum of squared normal variates and is therefore distributed as Chi Square – χ^2

$$C.I. = \frac{(n-1)S^2}{\chi_{1-\alpha/2, n-1}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{\alpha/2, n-1}^2}$$

where:

n = the sample size

S^2 = the sample variance

$\frac{(n-1)S^2}{\chi_{1-\alpha/2, n-1}^2}$ & $\frac{(n-1)S^2}{\chi_{\alpha/2, n-1}^2}$ = the χ^2 distribution values for the desired confidence level α and for $n-1$

Hypothesis testing

Binary true/false validity test of hypotheses about population parameters

We have a hypothesis we would like to test.

The approach is not to use the data to accept this hypothesis as true but to reject the null hypothesis that a value could have arisen by chance

If data are found to be inconsistent with the null hypothesis we reject the null hypothesis in favor of the alternate hypothesis

Hypothesis testing

Use a suitable test statistic $T(X)$ calculated from a sample e.g Z score

Set up a reasonable null hypothesis.

Specify a level of significance α – probability that the null hypothesis will be rejected even if it is really true – type 1 error.

Use H_0 to calculate the sampling distribution for the test statistic $T(X)$.

Calculate the p value $p = \Pr\{|T| \geq t\}$ for the observed sample value t of the test statistic.

Reject H_0 if the p value is less than the level of significance, $p < \alpha$

Hypothesis testing

The null hypothesis can postulate (suggest) that two samples are drawn from the same population, so that the variance and shape of the distributions are equal, as well as the means.

Other types of null hypotheses may be, for example, that:

- values in samples from a given population can be modeled using a certain family of statistical distributions.
- the variability of data in different groups is the same, although they may be centered around different values.

Hypothesis testing

The level of significance defines the rejection region (critical region) in the tails of the sampling distribution of the test statistic

If the observed value of the test statistic lies in the rejection region, the p-value is less than α and we reject the null hypothesis

Hypothesis test example

Assume we know a normally distributed population mean of 170cm and a population standard deviation of 30cm. Sample size = 11

1. State the null hypothesis and the alternative

$$H_0 : \mu = \mu_0 \quad \text{where } \mu_0 = 170\text{cm}$$

$$H_1 : \mu \neq \mu_0$$

2. Specify level of significance – 0.05

Based on the null hypothesis the sample mean is distributed as:

$$\bar{X} \sim N(\mu_0, \sigma_0^2/n) \quad Z = (\bar{X} - \mu_0)/(\sigma_0/\sqrt{n}) \sim N(0, 1)$$

Hypothesis test example

Hence test statistic $Z = (\bar{X} - \mu_0)/(\sigma_0/\sqrt{n}) \sim N(0, 1)$

$$\bar{X} = 174.3$$

$$Z = (174.3 - 170)/30/\sqrt{11}$$

$$Z = .4753$$

$$p\text{-value} = .6455$$

P-value is much greater than the significance value so we can not reject the null hypothesis

Hypothesis testing

Possible situations that can arise in hypothesis testing

	H_0 true	H_1 true
$p > \alpha$ don't reject H_0	Correct non-rejection probability $1 - \alpha$	Missed rejection (Type II error) probability β
$p \leq \alpha$ reject H_0	False rejection (Type I error) probability α	Correct rejection probability $1 - \beta$

Hypothesis testing

One-tailed tests – one critical region

$$\left[-\infty, \bar{x} + t_{n-1, 1-\alpha} \frac{s}{\sqrt{n}} \right] \quad \text{One-sided upper confidence interval for t distribution}$$

$$\left[\bar{x} - t_{n-1, 1-\alpha} \frac{s}{\sqrt{n}}, \infty \right] \quad \text{One-sided lower confidence interval for t distribution}$$

Hypothesis testing

T test for non-zero correlation

The significance (probability) of the correlation coefficient is determined from the t-statistic. The probability of the t-statistic indicates whether the observed correlation coefficient occurred by chance if the true correlation is zero. In other words, it asks if the correlation is significantly different than zero.

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

$$df = (n - 2)$$

Whenever we perform a significance test, it involves comparing a test value that we have calculated to some critical value for the statistic. It doesn't matter what type of statistic we are calculating (e.g., a t-statistic, a chi-square statistic, an F-statistic, etc.), the procedure to test for significance is the same.

1. Decide on the critical alpha level you will use (i.e., the error rate you are willing to accept).
2. Conduct the research.
3. Calculate the statistic.
4. Compare the statistic to a critical value obtained from a table.

If your statistic is higher than the critical value from the table:

- Your finding is significant.
- You reject the null hypothesis.
- The probability is small that the difference or relationship happened by chance, and p is less than the critical alpha level ($p < \alpha$).

If your statistic is lower than the critical value from the table:

- Your finding is not significant.
- You fail to reject the null hypothesis.
- The probability is high that the difference or relationship happened by chance, and p is greater than the critical alpha level ($p > \alpha$).

Modern computer software can calculate exact probabilities for most test statistics. If you have an exact probability from computer software, simply compare it to your critical alpha level. If the exact probability is less than the critical alpha level, your finding is significant, and if the exact probability is greater than your critical alpha level, your finding is not significant. Using a table is not necessary when you have the exact probability for a statistic.