

# Introduction to Geostatistics

## Confidence intervals II: confidence intervals for differences, and in general.

Edzer J. Pebesma

edzer.pebesma@uni-muenster.de  
Institute for Geoinformatics (ifgi)  
University of Münster

summer semester 2007/8,  
May 25, 2009



## Point estimation vs interval estimation

- ▶ Point estimation is e.g. giving  $\bar{X}$  as an estimate of  $\mu$
- ▶ Obviously, we try always to give the “best” point estimate
- ▶ “best” usually has some mathematical connotation: least squares, minimum variance, best linear, maximum likelihood, maximum a-posteriori probability, ...
- ▶ A more complete picture is given by the *interval estimate*, where we give the **range of likely values** for the target parameter (e.g.  $\mu$ ), given sampling error
- ▶ this is usually done with a confidence interval that has a certain probability coverage (e.g. 95%)
- ▶ probability refers to sampling error/repeated sampling, not to the population parameter (such as  $\mu$ )



## Confidence intervals, $\sigma$ known

We saw that

$$Pr(\bar{X} - 1.96SE < \mu < \bar{X} + 1.96SE) = 0.95$$

and we can call this a **95% confidence interval**.

The essence is that we have limited knowledge about  $\mu$ , and this is what we can say about it, based on sampling data.

Other probabilities can also be obtained. Let  $\alpha$  be the probability that the confidence interval does *not* cover the true value, in this case 0.05.

$z_{\alpha/2}$  is the value of the standard normal curve below which  $\alpha/2$  probability lies. Then we obtain a confidence interval with  $1 - \alpha$  probability coverage by

$$[\bar{X} + z_{\alpha/2}SE, \bar{X} + z_{1-\alpha/2}SE]$$

(Note that  $z_{\alpha/2}$  is negative.)

Values for  $\alpha$ :

- ▶  $\alpha$  should be small, not larger than .1 for the word "confidence" to make sense
- ▶ Other values might be 0.1, 0.01, 0.001



## Confidence intervals, $\sigma$ known – example

A 99% confidence interval for Length, assuming  $\sigma = 11$ :

```
> load("students.RData")
> attach(students)
> m = mean(Length)
> sd = 11
> se = sd/sqrt(length(Length))
> alpha = 0.01
> c(m + qnorm(alpha/2) * se, m + qnorm(1 - alpha/2) * se)
[1] 175.7123 180.3548

> alpha = 0.05
> c(m + qnorm(alpha/2) * se, m + qnorm(1 - alpha/2) * se)
[1] 176.2673 179.7998

> alpha = 0.1
> c(m + qnorm(alpha/2) * se, m + qnorm(1 - alpha/2) * se)
[1] 176.5513 179.5158
```



## Confidence intervals, $\sigma$ unknown

What to do if  $\sigma$  is not known (and in real life, it isn't)?

We know that if  $n$  is large, we can estimate  $\sigma$  quite well with the sample standard deviation  $s$ . If however  $n$  is small, the approximation is worse.

We need a distribution that is like the normal distribution, but wider for smaller  $n$ . This is what the ***t*-distribution** does.

```
> sd = sqrt(var(Length))
> n = length(Length)
> se = sd/sqrt(n)
> alpha = 0.05
> c(m + qnorm(alpha/2) * se, m + qnorm(1 - alpha/2) * se)

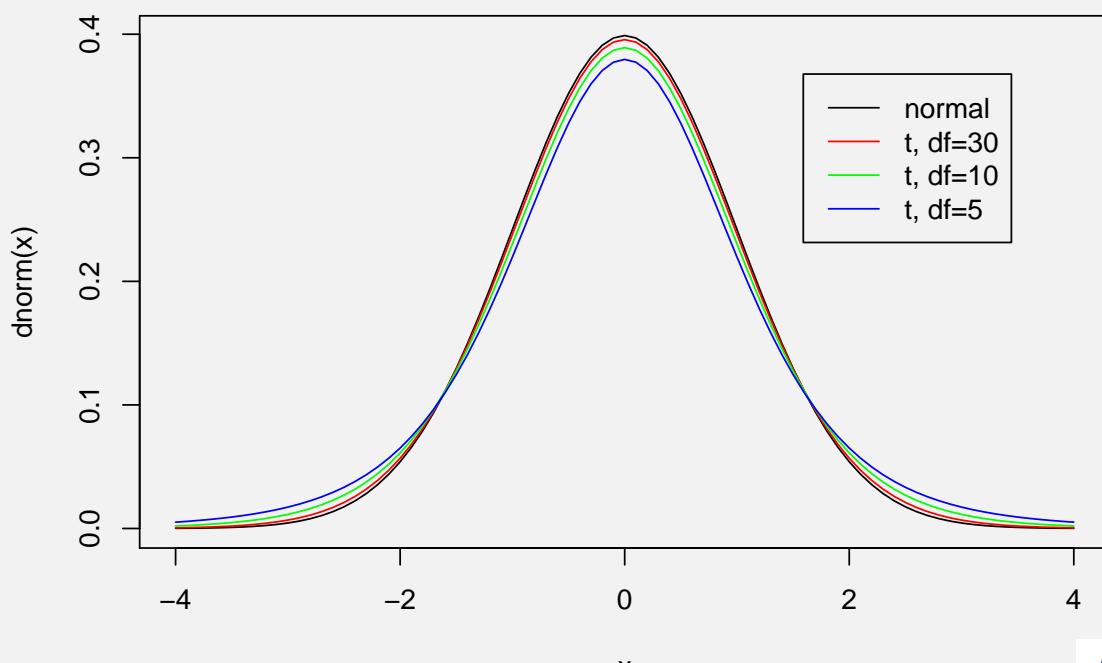
[1] 176.2752 179.7919

> c(m + qt(alpha/2, n - 1) * se, m + qt(1 - alpha/2, n -
+      1) * se)

[1] 176.2607 179.8064
```



## t-distribution



## small sample size:

```
> L10 = Length[1:10]
> m = mean(L10)
> se = sqrt(var(L10)/10)
> c(m + qnorm(alpha/2) * se, m + qnorm(1 - alpha/2) * se)
[1] 159.7252 162.8748

> c(m + qt(alpha/2, 9) * se, m + qt(1 - alpha/2, 9) * se)
[1] 159.4824 163.1176

> L5 = Length[1:5]
> m = mean(L5)
> se = sqrt(var(L5)/5)
> c(m + qnorm(alpha/2) * se, m + qnorm(1 - alpha/2) * se)
[1] 158.4666 159.9334

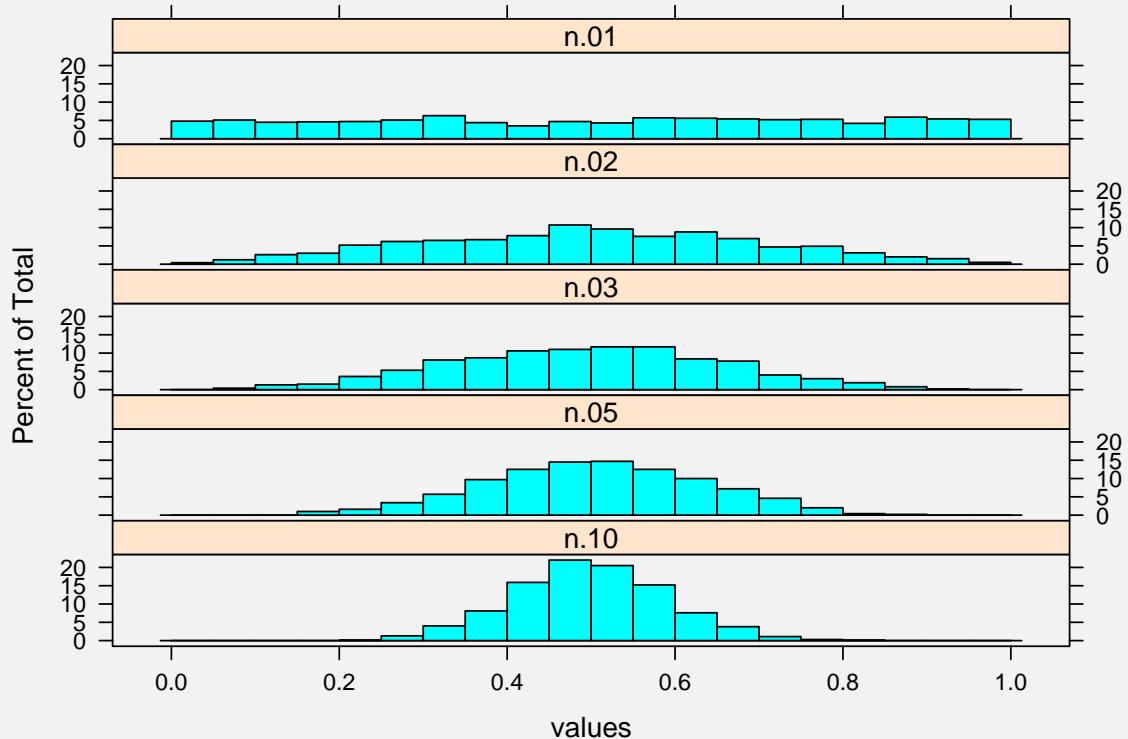
> c(m + qt(alpha/2, 4) * se, m + qt(1 - alpha/2, 4) * se)
[1] 158.1611 160.2389
```



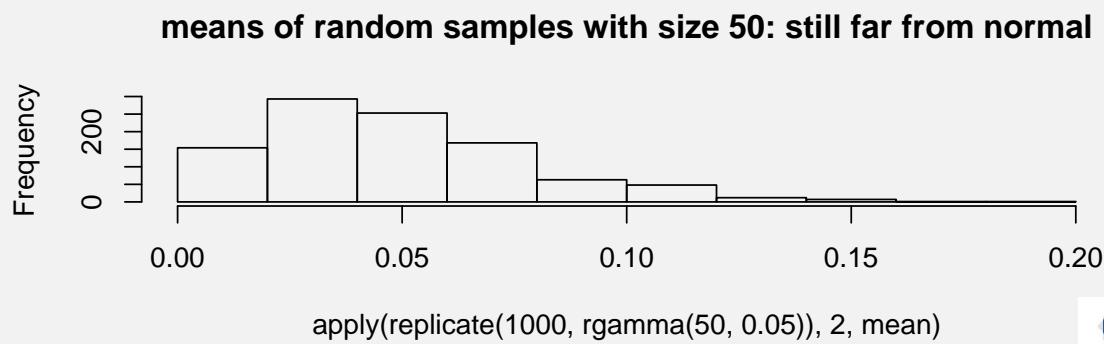
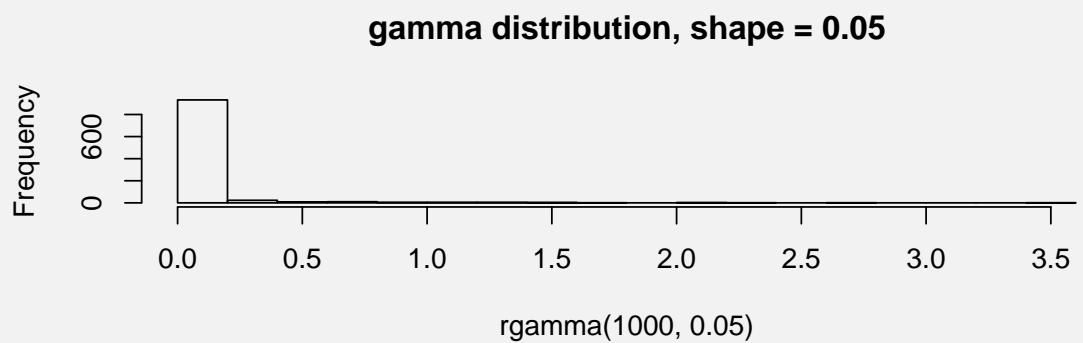
## The normal assumption

- ▶ When computing confidence intervals based on the normal distribution ( $\sigma$  known) or  $t$ -distribution ( $\sigma$  unknown) we assume normality. But normality of what?
- ▶ **NOT** of the data,  $X_i$ , but
- ▶ of the estimation error of the mean,  $\bar{X} - \mu$
- ▶ When is this assumption justified?
  1. when the data are (close to) normally distributed **OR**
  2. **when the sample size is large enough**
- ▶ when is a sample large enough? (usually:  $n > 30$ )





An example where it does not work out:



# Why does this normality thing work?

## The central limit theorem:

Loosely, this theorem states that if we take a sum of  $n$  independent random variables **with an arbitrary distribution**,

$$Y = \sum_{i=1}^n X_i$$

then, when  $n$  grows larger, then the distribution of  $Y$  will converge to a normal distribution. As the mean is also a sum, this applies to sample means. How fast is the convergence?



## CI for the difference in means; independent samples

Suppose we have two samples, and are interested in the difference in their means. We can now form a confidence interval for  $\mu_1 - \mu_2$ . What is the standard error for  $\bar{X}_1 - \bar{X}_2$ ? Suppose  $\sigma_1 = \sigma_2$ , then

$$SE = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left[ \frac{1}{n_1} + \frac{1}{n_2} \right]}$$

and the 95% confidence interval is

$$Pr((\bar{X}_1 - \bar{X}_2) - t_{df, \alpha} SE \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t_{df, \alpha} SE) = .95$$

The usual interest lies in whether this interval contains zero.



## CI for the difference in means; independent samples

```
> t.test(Length ~ Gender, var.equal = TRUE)

  Two Sample t-test

data: Length by Gender
t = -11.07, df = 147, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-17.84502 -12.43874
sample estimates:
mean in group female   mean in group male
          168.6842           183.8261
```



## CI for the difference in means; paired samples

Paired samples: a single object has been measured twice (usually at two moments, or "before" and "after" treatment)

obj	$t_1$	$t_2$
1	13.5	12.7
2	15.3	15.1
3	7.5	6.6
4	10.3	8.5
5	8.7	8.0

```
> x1 = c(13.5, 15.3, 7.5, 10.3, 8.7)
> x2 = c(12.7, 15.1, 6.6, 8.5, 8)
> x1 - x2

[1] 0.8 0.2 0.9 1.8 0.7
```



```

> t.test(x1, x2, var.equal = TRUE)
  Two Sample t-test

data: x1 and x2
t = 0.4066, df = 8, p-value = 0.695
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-4.111314 5.871314
sample estimates:
mean of x mean of y
11.06      10.18

> t.test(x1 - x2)

  One Sample t-test

data: x1 - x2
t = 3.3896, df = 4, p-value = 0.02754
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
0.1591929 1.6008071
sample estimates:
mean of x
0.88

```



## CI for (difference in) proportions

Proportions: use figure on page 274 (W&W) Large sample approximation:

$$P \pm 1.96 \sqrt{\frac{\pi(1 - \pi)}{n}}$$

by substituting  $P$  for  $\pi$  (for a conservative interval, i.e. worst case, substitute 0.5 for  $\pi$ ).

Difference in proportions, large sample approximation:

$$\Pr((P_1 - P_2) - 1.96\text{SE} \leq \pi_1 - \pi_2 \leq (P_1 - P_2) + 1.96\text{SE}) \approx .95$$

$$\text{with SE} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$



## Ratio's of variances: F distribution

- ▶ Suppose we have two samples, and are interested whether they come from two populations having different variances, i.e.  $\sigma_1 \neq \sigma_2$ . Let sample 1 be the group with the larger variance. The F distribution describes the ratio of two sample variances under  $H_0 : \sigma_1 = \sigma_2$ .
- ▶ Under the hypothesis that  $\sigma_1 = \sigma_2$ , the ratio  $\frac{s_1^2}{s_2^2}$  follows the F distribution with  $n_1$  and  $n_2$  degrees of freedom.
- ▶ Suppose that  $s_1^2 = 9$ ,  $s_2^2 = 3$   $n_1 = 20$ ,  $n_2 = 30$ , so the sample variance ratio is  $9/3=3$ .



```
> qf(0.95, 20, 30)
[1] 1.931653

> v1 = var(Length[Gender == "male"])
> v2 = var(Length[Gender == "female"])
> v1
[1] 42.51887

> v2
[1] 103.7556

> v2/v1
[1] 2.440226

> qf(0.95, length(Length[Gender == "female"]),
+      length(Length[Gender == "male"]))
[1] 1.468575
```



```
> t.test(Length ~ Gender, var.equal = TRUE)
  Two Sample t-test

data: Length by Gender
t = -11.07, df = 147, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-17.84502 -12.43874
sample estimates:
mean in group female  mean in group male
168.6842              183.8261

> t.test(Length ~ Gender)

  Welch Two Sample t-test

data: Length by Gender
t = -10.0226, df = 84.687, p-value = 4.809e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-18.14586 -12.13789
sample estimates:
mean in group female  mean in group male
168.6842              183.8261
```

