Introduction to Geostatistics

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

Edzer Pebesma

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

June 1, 2010

The normal assumption

- When computing confidence intervals based on the normal distribution (σ known) or t-distribution (σ 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?
    - when the data are (close to) normally distributed OR
- when is a sample large enough? (usually: $n > 30$)

The normal assumption

- When computing confidence intervals based on the normal distribution (σ known) or t-distribution (σ 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: $n > 30$)

The normal assumption

- When computing confidence intervals based on the normal distribution (σ known) or t-distribution (σ 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: $n > 30$)

The normal assumption

- When computing confidence intervals based on the normal distribution (σ known) or t-distribution (σ 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: $n > 30$)

The normal assumption

- When computing confidence intervals based on the normal distribution (σ known) or t-distribution (σ 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: $n > 30$)

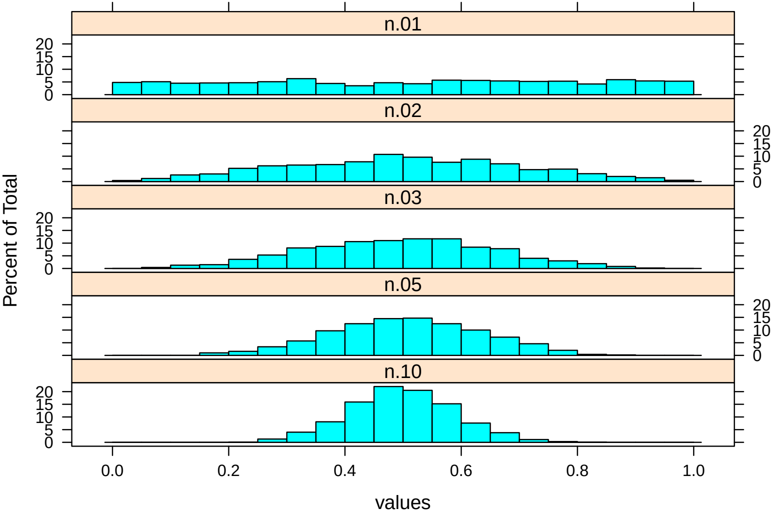
The normal assumption

- When computing confidence intervals based on the normal distribution (σ known) or t-distribution (σ 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: $n > 30$)

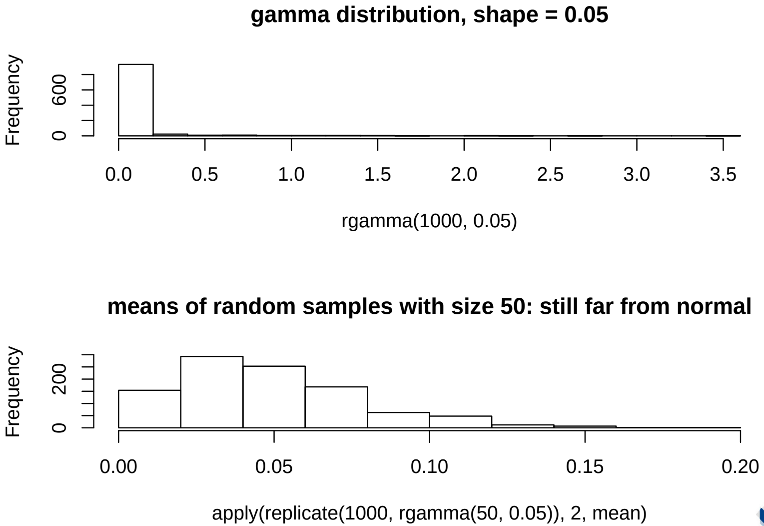
The normal assumption

- When computing confidence intervals based on the normal distribution (σ known) or t-distribution (σ 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: $n > 30$)

ifgi


n.01
n.02
n.03
n.05
n.10
Percent of Total
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0
values

An example where it does not work out:



ifgi

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?

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?

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 for a confidence interval for $\mu_1 - \mu_2$ What is the standard eror for $\bar{X}_1 - \bar{X}_2$? Suppose $\sigma_1 = \sigma_2$, then

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

and the 95% confidence interval is

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

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 = -13.3724, df = 245, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.25146 -11.33533
sample estimates:
mean in group female   mean in group male
            169.8495             183.1429
```

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 π (for a conservative interval, i.e. worst case, substitute 0.5 for π).
Difference in proportions, large sample approximation:

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

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

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.

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.

ifgi

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] NA

> v2

[1] NA

> v2/v1

[1] NA

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

[1] 1.347627
```

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

        Two Sample t-test

data:  Length by Gender
t = -13.3724, df = 245, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.25146 -11.33533
sample estimates:
mean in group female   mean in group male
            169.8495             183.1429

> t.test(Length ~ Gender)

        Welch Two Sample t-test

data:  Length by Gender
t = -12.3266, df = 148.535, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.42444 -11.16235
sample estimates:
mean in group female   mean in group male
            169.8495             183.1429
```