

Quantitative Methods in Geography

Geo 340

William D. McCoy

Department of Geosciences
University of Massachusetts, Amherst

Autumn, 2007

Parameter estimation

In statistical inference we make estimates of the parameters of a population based on samples. We can use a sample to help make a point estimate of the value of a population parameter.

Unfortunately, with just a point estimate we don't know how close we might be to the true value of the population parameter. But we can also use a sample to construct an interval estimate, *i.e.* an interval within which we can have a specified level of confidence that the true value of the population parameter lies.

We will first deal with making point estimates of some useful population parameters, and then we will discuss building confidence intervals.

Point estimates and confidence intervals

- ▶ point estimates of population parameters
 - ▶ unbiased estimators
 - ▶ minimum variance estimators
 - ▶ estimates of the population mean
 - ▶ estimates of the population variance
 - ▶ estimates of the mean of a proportion
- ▶ confidence intervals
 - ▶ confidence interval for the mean of a population
 - ▶ confidence interval for the mean when σ is unknown

Unbiased point estimators

A desirable property of a point estimator is that it be unbiased. An unbiased estimator has a sampling distribution expected value that is equal to the value of the population parameter:

$$E(\hat{\theta}) = \theta$$

Bias is measured as the mean error:

$$BIAS = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$$

Minimum variance estimators

Another desirable property of a point estimator is that it has low variance. An estimator having a sampling distribution with a small spread is preferable to one with a large spread. So we seek an estimator that has the smallest spread (minimum variance) as well as being unbiased. However, sometimes we might prefer a slightly biased estimator if it had much less variance than an unbiased alternative estimator. The variance of a sampling distribution is defined as:

$$VARIANCE = E[\hat{\theta} - E(\hat{\theta})]^2$$

Efficient estimators

The most efficient estimator is one that has a sampling distribution with low, or no, bias and low variance. The mean squared error (MSE) is a measure of the efficiency of an estimator. The lower the MSE (or its square root), the more efficient, and the more desirable, is the estimator.

$$MSE = E(\hat{\theta} - \theta)^2 =$$

$$[E(\hat{\theta}) - \theta]^2 + E[\hat{\theta} - E(\hat{\theta})]^2 =$$

$$BIAS^2 + VARIANCE$$

\sqrt{MSE} is known as the root mean squared error or RMSE.

Estimators of the population mean

The minimum variance, unbiased estimator (MVUE) and most efficient (lowest MSE) estimator of the population mean of a normal distribution is the sample mean, $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. The sample median also has an unbiased sampling distribution, but has greater variance. Likewise, trimmed means are unbiased estimators of the population means of symmetric distributions, and can be more efficient than the sample mean if the distribution has heavier tails than a normal distribution. Another very important property of the sample mean is that it has a known sampling distribution (see central limit theorem). Therefore, it is commonly used in statistical inference.

Estimator of the population variance

The minimum variance, unbiased estimator (MVUE) of the population variance is the sample variance,

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

S is not necessarily an unbiased estimator of σ , but its sampling distribution has low bias and so it is most often used as an estimator of σ .

Estimator of a population proportion

The best unbiased estimator for p , the probability of success in a Bernoulli trial and a parameter in binomial distributions, is $\hat{p} = \frac{X}{n}$, or simply the number of successes divided by the number of trials.

We can rearrange that equation to show that our sample mean should fall within some interval centered on the population mean ($\mu = \mu_{\bar{X}}$). We can calculate that interval to any desired level of confidence, $1 - \alpha$, where $\alpha/2$ is the area in each tail of the distribution.

$$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha$$

$$P(-z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$$P(\mu - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Interval estimation

From our discussion the central limit theorem and what it says about the sampling distribution of sample means taken from a normal distribution, we know that

$$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu_{\bar{X}}}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha$$

That is, $\frac{\bar{X} - \mu_{\bar{X}}}{\sigma/\sqrt{n}}$ is normally distributed with a mean of 0 and a standard deviation of 1. Since we know that distribution, we can calculate the quantiles $-z_{\alpha/2}$ and $z_{\alpha/2}$ between which we have $1 - \alpha$ of the area under the normal curve.

So, for example, if we want the .95 (95%) confidence interval, we have

$$P(\mu - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 1.96 \cdot \frac{\sigma}{\sqrt{n}}) = 0.95.$$

We can say that 95% of the means of all samples of size n will lie within the interval between $\mu - 1.96\sigma_{\bar{X}}$ and $\mu + 1.96\sigma_{\bar{X}}$.

Confidence interval for the mean

We can also rearrange that equation to give an expression of the interval that will contain the true population mean, μ , with some desired level of confidence, $1 - \alpha$.

$$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha$$

$$P(-z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$$P(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Confidence interval for the mean (σ known)

So, for example, if we want the .95 (95%) confidence interval, we have

$$P(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}) = 0.95.$$

We can say that 95% of the confidence intervals that we construct based on different samples of size n will contain the true population mean, μ .

So, in general, we can say that when σ is known, the $1 - \alpha$ confidence interval for μ is

$$\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

Confidence interval for the mean (σ unknown)

However, when σ is not known, which is usually the case, we have to base our estimate of σ on S , the standard deviation of the sample, which is itself a random variable. So now our confidence intervals must be based on the t -distribution, which is broader than the standard normal distribution when n is small. The $1 - \alpha$ confidence interval for μ is then

$$\bar{X} \pm t_{\alpha/2, n-1} \cdot \frac{S}{\sqrt{n}}.$$

Remember, however, that when n is small the central limit theorem does not apply to non-normal populations, so we need to be sure that our underlying population is at least approximately normally distributed for our confidence interval to be valid.

Using **R** to find a confidence interval for the mean

The **R** function `t.test()` can be used to estimate an interval for the mean of a population with a specified level of confidence ($1 - \alpha$). To find a 95% confidence interval for the population mean of weights of babies born to mothers who never smoked, we can first subset the data we want from the dataset `babies`:

```
> library(UsingR)
> babiesNS <- subset(babies, smoke == 0,
+   select = wt, drop = TRUE)
```

Confidence intervals with summary data

Sometimes you are given summary data from a sample, but don't have the sample data itself. As long as you have the sample mean, standard deviation, and sample size, you can find a confidence interval for the mean, but you can't use the `t.test()` function. Suppose you know that a random sample of heights of size 25 was taken from an approximately normally distributed population of heights of 20-year-old males. The mean was 68.5 inches and the standard deviation was 3.5 inches. To find the 95% confidence interval, you can calculate:

```
> xbar <- 68.5
> sem <- 3.5/sqrt(25)
> tstar <- qt(0.975, 24)
> c(xbar - tstar * sem, xbar + tstar * sem)

[1] 67.05527 69.94473
```

Confidence intervals for proportions, p

So far we have only discussed confidence intervals for the mean of a population. If we have data on successes from a presumed binomial distribution or summarized data of a proportion, we can construct a confidence interval using `prop.test()` or `binom.test()`.

Suppose we have a random sample of 150 likely voters who were asked if they intended to vote for gubernatorial candidate Jane Smith. From the sample we find that 93 (62%) of the voters say they will vote for her. What is the 95% confidence interval for the proportion of likely voters who will vote for Jane Smith?

```
> t.test(babiesNS)
One Sample t-test

data: babiesNS
t = 167.3698, df = 543, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 121.3366 124.2186
sample estimates:
mean of x
 122.7776
```

A 99% confidence interval would necessarily need to be wider. To see what it is, specify the `conf.level` argument:

```
> t.test(babiesNS, conf.level = 0.99)
One Sample t-test

data: babiesNS
t = 167.3698, df = 543, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 120.8814 124.6738
sample estimates:
mean of x
 122.7776
```

The first argument to `prop.test()` is the number of "successes" and the second argument is the sample size. To specify confidence level, use the `conf.level` argument. The default confidence level is 0.95.

```
> prop.test(93, 150)
```

```
1-sample proportions test with continuity  
correction
```

```
data: 93 out of 150, null probability 0.5
```

```
X-squared = 8.1667, df = 1, p-value =
```

```
0.004267
```

```
alternative hypothesis: true p is not equal to 0.5
```

```
95 percent confidence interval:
```

```
0.5368693 0.6968841
```

```
sample estimates:
```

```
p
```

```
0.62
```

Use `binom.test()` in the same way:

```
> binom.test(93, 150)
```

```
Exact binomial test
```

```
data: 93 and 150
```

```
number of successes = 93, number of trials =
```

```
150, p-value = 0.004113
```

```
alternative hypothesis: true probability of success is not
```

```
95 percent confidence interval:
```

```
0.5372369 0.6979228
```

```
sample estimates:
```

```
probability of success
```

```
0.62
```