

BIOS 600: Principles of Statistical Inference

Interval Estimation

Fall 2012

Reading

- ▶ Pagano and Gauvreau, Chapter 9

Arsenic in Rice!

On September 19, 2012, the FDA released data on arsenic levels in rice, which is contained in the file [arsenicrice.dta](#) on the Sakai website. A friend who is making Rice Krispy treats hears about the study on NPR and calls you to ask whether they should be consumed or not (you are taking a public health class!).



Forms of Statistical Inference

- ▶ *Point estimation*: estimating an unknown parameter using a single number calculated from the sample data
 - ▶ In our sample of North Carolina children, we estimate that 36% of children receive preventive dental services before age 3
- ▶ *Interval estimation*: estimating an unknown parameter using an interval or range of values that is likely to cover the true population value
 - ▶ We estimate that 33-39% of children in North Carolina receive preventive dental services before age 3
- ▶ *Hypothesis testing*: checking whether sample data provide evidence against some claim made about the population
 - ▶ We evaluated the hypothesis that family socioeconomic status was unrelated to receipt of preventive dental services before age 3. In our sample the proportion of low-income children receiving preventive dental services by this age was roughly half that of high-income children, providing evidence against this hypothesis.

Point Estimation

Point estimation involves making a single (hopefully 'best' in some sense) 'guess' about the value of a parameter in the population. Commonly used point estimates include the sample mean, sample median, or sample proportion.

- ▶ A September 2012 Gallup poll estimates that if the election were held today, 48% would vote for Obama and 46% would vote for Romney. If the election were held today, do we expect *exactly* those figures to support each candidate?
- ▶ We would not be surprised to have 48.1% vote for Obama, or 46.3% vote for Romney, given that poll result, right?
- ▶ What about 48.5%? 49%? 53%? 42%? We are really interested in knowing the *margin of error* of the poll so that we have a good idea about a plausible range of values.
- ▶ The idea behind *interval estimation* is to convey information about the amount of error contained in an estimate.

Interval Estimates Matter!



Which quote do you prefer?

- ▶ \$1200
- ▶ Probably around \$700, but could be as high as \$1400 or as low as \$500, depending on how long it takes

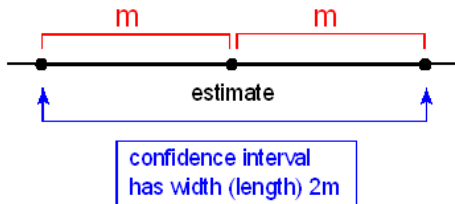
What is a Confidence Interval?

A *confidence interval* provides a range of reasonable values that are intended to contain the parameter of interest with a certain degree of confidence. It often takes the form

$$\text{point estimate} \pm \text{margin of error}$$

and is written

(point estimate $-$ margin of error, point estimate $+$ margin of error).



Caveat

For illustration, we start by assuming σ is known.

- ▶ When is σ known?
- ▶ Almost never!
- ▶ However, it's easier to understand if we assume that to start.
By the end of the class, we'll get rid of this assumption.

Two-Sided Confidence Intervals

We use what we learned about the sampling distribution of the mean, and about the normal distribution, to construct a confidence interval for μ when σ is known. Given a random variable X with mean μ and standard deviation σ , we know from the central limit theorem that

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

has a standard normal distribution if X is normal itself and an approximate standard normal distribution if X is not normal but n is large enough.

For a standard normal random variable, recall that 95% of the observations lie between -1.96 and 1.96, so that

$$Pr(-1.96 \leq Z \leq 1.96) = 0.95.$$

Two-sided confidence intervals

Now substitute for $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ and use algebra to rewrite as

$$Pr(-1.96 \leq Z \leq 1.96) = 0.95$$

$$Pr\left(-1.96 \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96\right) = 0.95$$

$$Pr\left(-1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$Pr\left(-\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$Pr\left(\bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$Pr\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Two-sided Confidence Intervals

So our 95% confidence interval (95% CI) is given by

$$\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right).$$

Be careful interpreting this interval.

- ▶ If we select a number m of random samples from the population and use them calculate m different confidence intervals for μ , then approximately 95% of the intervals would cover the true population mean μ , and 5% would not.
- ▶ Sometimes people interpret the interval by saying that they are “95% confident” that the interval covers the true mean. This is ok as long as you philosophically can feel “95% confident” about something (to me, confident is an absolute adjective, like ‘dead’ or ‘perfect,’ so I avoid this usage).

Two-sided Confidence Intervals

So our 95% confidence interval (95% CI) is given by

$$\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right).$$

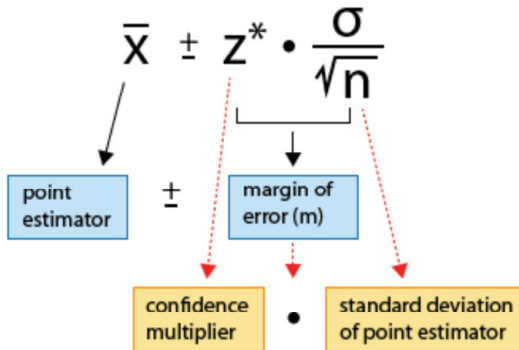
Be careful interpreting this interval.

- ▶ We don't know whether any one interval is in the 95% that would cover the mean, or the 5% that would not. (Sorry. I know that's what everyone wants. This is one reason Bayesian inference is nice!)
- ▶ **WRONG:** There is a 95% chance that μ lies in the interval (μ is fixed, and either it is in the interval, or it is not)

Confidence Interval Interpretation

Confidence interval simulation

Confidence Intervals



Confidence Intervals with Other Coverage Probabilities

While 95% confidence intervals are the most common, it is simple to generate other intervals, for example 99% intervals. The only change is that you replace the z-score 1.96 (cuts of 2.5% in each tail) with the z-score that cuts off the appropriate amount (0.5% in each tail for a 99% interval). So the general formula is

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

for a $100\% \times (1 - \alpha)$ confidence interval. Here the notation $z_{\frac{\alpha}{2}}$ indicates the z score that cuts off the upper $100 \times \frac{\alpha}{2}\%$ of the distribution. So if $\alpha = 0.05$ you have a 95% interval using $z_{0.025} = 1.96$.

99% Confidence Interval

For a 99% interval, we need the z-value that cuts off the top 0.5% or 0.005 of the distribution, which is _____.

Normal Table

TABLE A.3

Areas in the upper tail of the standard normal distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.496	0.492	0.488	0.484	0.480	0.476	0.472	0.468	0.464
0.1	0.460	0.456	0.452	0.448	0.444	0.440	0.436	0.433	0.429	0.425
0.2	0.421	0.417	0.413	0.409	0.405	0.401	0.397	0.394	0.390	0.386
0.3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348
0.4	0.345	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312
0.5	0.309	0.305	0.302	0.298	0.295	0.291	0.288	0.284	0.281	0.278
0.6	0.274	0.271	0.268	0.264	0.261	0.258	0.255	0.251	0.248	0.245
0.7	0.242	0.239	0.236	0.233	0.230	0.227	0.224	0.221	0.218	0.215
0.8	0.212	0.209	0.206	0.203	0.200	0.198	0.195	0.192	0.189	0.187
0.9	0.184	0.181	0.179	0.176	0.174	0.171	0.169	0.166	0.164	0.161
1.0	0.159	0.156	0.154	0.152	0.149	0.147	0.145	0.142	0.140	0.138
1.1	0.136	0.133	0.131	0.129	0.127	0.125	0.123	0.121	0.119	0.117
1.2	0.115	0.113	0.111	0.109	0.107	0.106	0.104	0.102	0.100	0.099
1.3	0.097	0.095	0.093	0.092	0.090	0.089	0.087	0.085	0.084	0.082
1.4	0.081	0.079	0.078	0.076	0.075	0.074	0.072	0.071	0.069	0.068
1.5	0.067	0.066	0.064	0.063	0.062	0.061	0.059	0.058	0.057	0.056
1.6	0.055	0.054	0.053	0.052	0.051	0.049	0.048	0.047	0.046	0.046
1.7	0.045	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.038	0.037
1.8	0.036	0.035	0.034	0.034	0.033	0.032	0.031	0.031	0.030	0.029
1.9	0.029	0.028	0.027	0.027	0.026	0.026	0.025	0.024	0.024	0.023
2.0	0.023	0.022	0.022	0.021	0.021	0.020	0.020	0.019	0.019	0.018
2.1	0.018	0.017	0.017	0.017	0.016	0.016	0.015	0.015	0.015	0.014
2.2	0.014	0.014	0.013	0.013	0.013	0.012	0.012	0.012	0.011	0.011
2.3	0.011	0.010	0.010	0.010	0.010	0.009	0.009	0.009	0.009	0.008
2.4	0.008	0.008	0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.006
2.5	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005
2.6	0.005	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
2.7	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
2.8	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2.9	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001
3.0	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.1	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.2	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3.4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

When can we use this CI?

The CI given by

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

is safe to use in the following circumstances when σ is known.

- ▶ X is normal (regardless of sample size)
- ▶ X is non-normal but the sample size is large

It is typically not safe to use this CI when the sample size is small and X is not a normal random variable.

How can we get a more narrow CI?

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

- ▶ Compromise on our level of confidence (e.g., a 90% interval)
 - ▶ Journals and other researchers will not like this strategy!
- ▶ Increase the sample size, n !

Did you get it?

Try these exercises for practice!

In a recent study of 500 randomly selected statistics students, they were asked the number of hours per week they spend studying for their statistics classes. The results were used to estimate the mean time for all statistics students with 90%, 95% and 99% confidence intervals. These were (not necessarily in the same order):

(7.5, 8.5) (7.6, 8.4) (7.7, 8.3)

One-sided Confidence Intervals

Sometimes, but not often, we want only an upper limit or a lower limit for the population mean. One example of this would be in a non-inferiority clinical trial for a generic version of a popular pharmaceutical product (we don't expect the generic drug to work better, but we do expect it to be just as good). Pagano Section 9.2 discusses construction of one-sided CI's.

What if σ is unknown?

The CI given by

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

cannot be computed if σ is unknown.

- ▶ Good news: back in Pagano Chapter 3 we learned how to calculate s , the sample standard deviation, as an estimate of σ
- ▶ Bad news: we never know σ , and if we replace σ with s , then
 - ▶ we can't use the central limit theorem
 - ▶ \bar{X} isn't exactly normal
 - ▶ so that using z scores no longer always works

What do you do when σ is unknown?



Really, I'm serious!

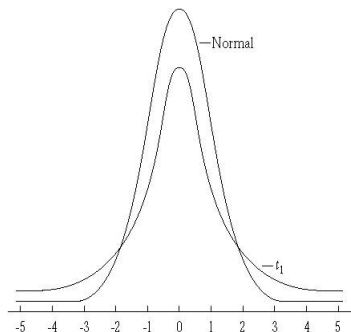
While working for Guinness Brewery in Dublin, William Sealy Gosset published a paper on the t distribution, which became known as Student's t distribution (he published under "Student" because the brewery didn't allow him to use his own name)



- ▶ He used the new distribution to determine how large a sample of persons to use in taste-testing beer!
- ▶ The t distribution is appropriate for constructing a confidence interval for the mean when we need to account for the additional variability due to estimating σ with s

Student's t distribution

- ▶ The t distribution looks a lot like the normal except that it has fatter tails
- ▶ The fatter tails lead to wider confidence intervals, which acknowledge our extra uncertainty because we had to estimate σ instead of using its true value
- ▶ As the sample size gets bigger, the t distribution looks more and more like the normal distribution



Student's t distribution

The t distribution has a property called *degrees of freedom*, abbreviated *df*. The degrees of freedom measure the amount of information available in the data to estimate σ and thus give us information about how reliable our estimate s is. The random variable

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

has a Student's t distribution with $n - 1$ degrees of freedom, represented using the notation t_{n-1} . (The df are $n - 1$ instead of n because we lose 1 df by estimating the sample mean \bar{x} .) For each possible df, there is a different t distribution, with the t distribution looking more like the normal as n gets large and s gets to be a better and better estimate of σ .

What if σ is unknown?

When σ is unknown, we use the CI given by

$$\left(\bar{X} - t_{n-1, \frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \quad \bar{X} + t_{n-1, \frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right)$$

TABLE A.4
Percentiles of the t distribution

df	Area in Upper Tail					
	0.10	0.05	0.025	0.01	0.005	0.0005
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
50	1.299	1.676	2.009	2.403	2.678	3.496
60	1.296	1.671	2.000	2.390	2.660	3.460
70	1.294	1.667	1.994	2.381	2.648	3.435

Case Study: Arsenic in Rice

Let's open the data set [arsenicrice.dta](#) and see how worried our friend should be.

Case Study: China Health and Nutrition Study

In these data, we do not know σ . Concentrating on the $n = 67$ subjects who are age 25 in 2009, we can quickly calculate a 95% confidence interval using the formula $\left(\bar{X} - t_{66,0.025} \frac{s}{\sqrt{67}}, \bar{X} + t_{66,0.025} \frac{s}{\sqrt{67}}\right)$ in Stata as follows, along with a 99% CI. First we need to create a BMI variable just for the 25 year olds.

```
. generate bmi25=bmi

.
. replace bmi25=. if age2009>=26
(7928 real changes made, 7928 to missing)

.
. replace bmi25=. if age2009<25
(1198 real changes made, 1198 to missing)

.
. summarize bmi25
```

Variable	Obs	Mean	Std. Dev.	Min	Max
bmi25	67	21.76736	3.472745	15.83899	31.03741

Case Study: China Health and Nutrition Study

Now we generate the CI's.

```
. ci bmi25
```

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]	
bmi25	67	21.76736	.4242634	20.92029	22.61443

```
. ci bmi25, level(99)
```

Variable	Obs	Mean	Std. Err.	[99% Conf. Interval]	
bmi25	67	21.76736	.4242634	20.64205	22.89267

So the 95% CI for BMI of 25 year olds is (20.92, 22.61) and the 99% CI is (20.64, 22.89).

Reading for Next Time

- ▶ Pagano and Gauvreau, Chapter 10
- ▶ Put It to the Test