**Module 7: Standard Errors and Confidence Intervals**

# Reporting Statistical Error

When we sample *x*1, *x*2, …, *xn* from a population and use  to estimate *µ*, it is extremely unlikely that  will be exactly the same as *µ*. Why?

Fortunately, the Central Limit Theorem (pp. 6-12) tells us that when the sample is *iid*



which means that  will be within  of *µ* about 95% of the time.

This expression, however, is not useful as given because in virtually every problem, σ is unknown.

# The Standard Error of

Powerful idea:

Since *s* estimates σ, we can use  to estimate .

The expression  is

Called the standard error of .

An estimate of the standard deviation of the sampling distribution of .

Often reported with  to indicate its precision.

Note

We are estimating the amount of sampling variation of  from the information in just the one observed sample *x*1, *x*2, …, *xn*.

# Example: Returns on Microsoft Stock

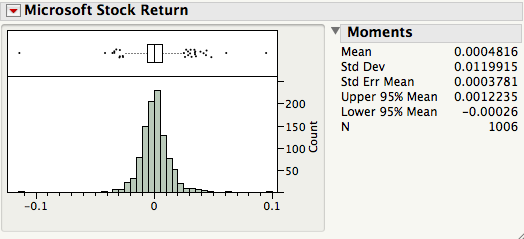
For the 2004-2007 Microsoft returns, we estimated *μ* by = .00048

and estimated σ by *s* = .0120 with *n* = 1,006 observations.

In this example, the standard error of is 0.012/√1006 = .00038.

What does this standard error convey?

The previous JMP output for the returns reports the standard error of .



# The Standard Error of the M&M’s Estimates

We saw in Module 6 that the standard deviation of the class estimates was

Bad news: we used the variation amount the many class samples to calculate this number. This is useless when we only have one sample.

Good news: an effective alternative is the standard error  and it can be calculated from a single sample!

Using one bag of M&M’s, we obtain[[1]](#footnote-1)

*s* = and  =

Is this close to our class findings?

# A 95% Confidence Interval for μ

A convenient way to convey the precision of a statistical estimate is to report a range of “probable” values of the parameter. This is done by reporting a *confidence interval*.

When *x*1, *x*2, …, *xn* is an iid sample from a population, then the interval spanned by the endpoints



is called a 95% confidence interval (CI) for the population mean *μ*.[[2]](#footnote-2)

The “95%” part of this name is known as the *confidence level* of the interval.

# Example

For the returns on Microsoft, =.00048, *s* = .0120, and *n* = 1,006

The resulting (approximate) 95% CI for *μ* is

(0.00048 – 2(0.0120/√1006), 0.00048 + 2(0.0120/√1006)) = (-.0003, .0012).

Key property: 95% CIs contain the population mean *μ* for 95% of samples.

To see why, recall from page 7-1 that



Hence, if we estimate σ by *s,*



Thus, approximately 95% of samples have the property that the sample mean  lies within two standard errors 2 of *μ*, or said differently, that *μ* is within the interval from  to .

# Exact Confidence Intervals For μ

If the data are a sample from a normal population, *x*1, *x*2, … *x*1 iid ~ N(*μ*, σ2), then the endpoints of exact confidence intervals for *μ* take the form



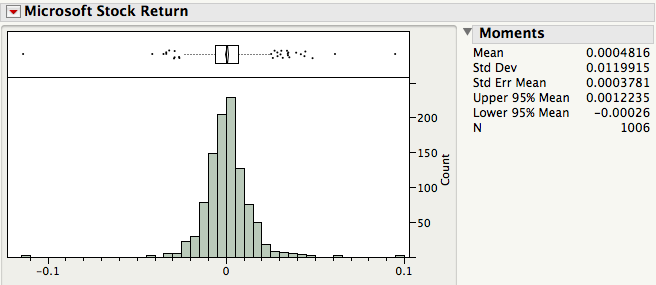
where *t* is a predetermined constant that depends on the sample size *n* and the desired confidence level (again, this is usually 95%).

For example, if *n* = 20, we can see the “price” of not knowing σ and having to estimate it from the data. The confidence interval is slightly longer.

*t* = 2.09 yields a 95% CI and *t* = 1.72 yields a 90% CI

JMP provides[[3]](#footnote-3) these exact 95% CI limits for *μ*.

Example: We can now explain the diamond that JMP shows by default in the boxplot. It’s the 95% CI for *μ*. For the 2004-2007 returns on Microsoft, we get this summary:



The 95% CI for *μ* here uses the exact procedure. How does it compare with the approximate interval (-.0003, .0012) on page 7-6?

# The "±2 Standard Error" Rule of Thumb

Unless *n* is small (≤ 30) and precise confidence is needed, the approximate 95% CI identified by  is fine.[[4]](#footnote-4)

# Confidence Intervals are not Prediction Intervals!

When the data are a sample from a normal population, *x*1, *x*2, …, *xn* iid ~ N(*μ*, σ2), an (approximate) 95% *prediction interval* [[5]](#footnote-5) for a future draw from this population is defined by the endpoints



What is the interpretation of this interval?

Careful!These are *not* the same as the (approximate) endpoints of the 95% CI for the population mean μ,



The prediction interval is wider because individual values are more variable than averages.

Example: Let’s revisit the 400 observations of motor shaft diameters in *ShaftXtr.JMP*.



A 95% CI for the mean of the shaft-making process is (814.90, 815.08).

Assuming the process remains in control[[6]](#footnote-6), an approximate 95% tolerance interval for the diameter of a future shaft is

(814.99 – 2(.93), 814.99 + 2(.93)) = (813.13, 816.85).

Which is wider: the confidence interval or tolerance interval? Why?

# The Effect of the Sample Size

The width of the 95% CI for the mean of the motor shaft process is (using all of the data) 815.1–814.9 = 0.2 (pp 7-10). What would happen to the width of the 95% CI for the mean if the sample size *n* were larger?

Using the ±2 standard error rule of thumb, the width of a 95% CI for the mean is

Width = 4 

Suppose we doubled the sample size from *n* to 2*n*. What would happen to the width of the interval?

The *margin of error* of an estimate such as reported in the media is conventionally half the width of the 95% confidence interval, namely 2.

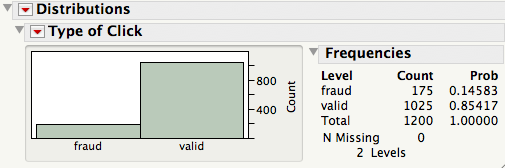
When estimating a population proportion such as in a political poll, the maximum margin of error for the 95% confidence interval is approximately . This can be used to choose the sample size *n* needed to guarantee a 95% level of accuracy for a desired margin of error.[[7]](#footnote-7)

# Click Fraud

A specialty retailer pays an on-line host for referrals obtained from on-line ads. The retailer pays the hosting site for each click on an ad that brings customers to its web site. Recently, however, the retailer suspects that many of these click-throughs have been generated by automated systems designed to imitate real customers. The on-line host has promised that no more than 10% of the clicks are imitations.

To learn more, the retailer hired a service to identify fraudulent clicks. In a sample of 1,200 clicks, the service identified 175 computer-generated fraudulent clicks. The file *Click\_Fraud.JMP* summarizes these counts.[[8]](#footnote-8)

The JMP summary[[9]](#footnote-9) for the nominal variable Type of Click gives



The sample proportion 14.58% is clearly larger than the 10% allowance promised by the hosting web site, but could this just be an accident, a fluke? Or does it indicate a more systematic problem in clicks that the retailer pays for? Before opening negotiations and bringing in the lawyers, the retailer would like to know if these results are compelling.

What does the sample reveal about the proportion of fraud in the population of clicks?

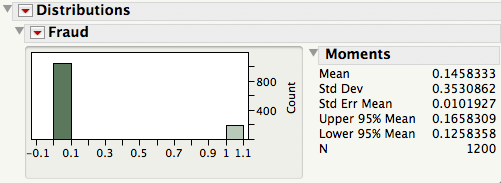
Assuming the survey has truly sampled the population of interest, the company would like to know a range of values for the population proportion of fraudulent clicks.

To answer this question with a confidence interval, we analyze the dichotomous variable Fraud which is a numerical encoding of the categorical variable:[[10]](#footnote-10)

Fraud = 0 if Type of Click = valid

Fraud = 1 if Type of Click = fraud

Because Fraud is numerical, the JMP summary provides moments and the 95% CI for the (unknown) population proportion.



Note that the sample proportion 14.58% is the mean  of Fraud.[[11]](#footnote-11)

What course of action would you recommend to managers of the retailer?

# Take-Away Review

The estimated standard error of  is  which can be from the sample that determines the value of . The expression uses the sample SD *s* to estimate the population SD σ.

To form a 95% confidence interval for the population mean, we can use the exact interval endpoints provided by JMP or the approximate “back of the envelope” interval



This form is typical of most of the CIs we will use in this course and Stat 621, namely

(estimated value) ± 2 *se*(estimated value)

# Next Module

If a value lies outside of the confidence interval, our analysis today suggests that “it’s not plausible” at some level of confidence. But once a value lies out of the interval, that’s about all we can say.

Statistical tests take this idea further and directly measure the reasonableness of a hypothesized population value.

1. An easily computed approximation to *s* in this context  which works since the average in the case of 0/1 data is the proportion of 1’s. See SF, page 370, for a careful description of situations in which averages are proportions. [↑](#footnote-ref-1)
2. Technically, this interval is an approximate 95% interval because we use 2 and estimate σ, the population SD, by *s*, the sample SD. An exact interval for samples from a normal population uses a percentile from a t-distribution in place of 2. See SF, section 15.2, page 357 and page 7.7 below. By varying the number of standard errors, we can obtain confidence intervals with different confidence levels. [↑](#footnote-ref-2)
3. Using Analyze > Distribution, it is listed in the Moments output. [↑](#footnote-ref-3)
4. The sample size for this rule to work does depend on how non-normal the data are. For instance, highly skewed samples may require a larger sample size for this interval to work. See the “Sample size condition” in SF, page 328. [↑](#footnote-ref-4)
5. In the context of statistical process control, such prediction intervals are sometimes called tolerance intervals. [↑](#footnote-ref-5)
6. Control charts monitor the stability of a process; see SF Chapter 14. [↑](#footnote-ref-6)
7. For more discussion of the margin of error and determining necessary sample sizes, see SF, section 15.5, page 364. [↑](#footnote-ref-7)
8. The data are represented compactly using a frequency column. Rather than enter, for example, 175 identical rows for the fraudulent clicks, the file has one row with value 175 in the column called “Freq”. This column tells JMP how many cases each row represents. [↑](#footnote-ref-8)
9. Using Analyze > Distribution, the output for a categorical variable takes this form. [↑](#footnote-ref-9)
10. 0/1 encodings of a categorical variable will be common in Stat 621. They’re known as dummy variables. [↑](#footnote-ref-10)
11. For more on the connection between averages, proportions, and 0/1 dummy variables, see “Behind the Math”, SF, p 370. [↑](#footnote-ref-11)