3.5: Statistics of error

Introduction

In this section, following the discussion about error in statistics, you'll find a justification for use of confidence intervals, how to calculate confidence intervals, both as an approximation and with an example of exact calculation, use of confidence interval to quantify accuracy, and conclude with a brief discussion of rounding and significant figures.

Statistics of error

An **error** in statistics means there was a difference between the measured value and the actual value for an object. Classical statistical approach developed a large body of calculated statistics, e.g., standard error of the mean, which allows the user to quantify how large the error of measurement is given assumptions about the distribution of the errors. Thus, classical statistics requires user to make assumptions about the error distribution, the subject of our Chapter 6. A critical issue to understand is that these methods assume large sample sizes are available; they are called **asymptotic statistics** or **large-sample statistics**; the properties of the statistical estimates are evaluated as sample size approaches infinity.

Jackknife sampling and **bootstrap sampling** are permutation approaches to working with data when the **Central Limit Theorem** — as sample size increases, the distribution of sample means will tend to a normal distribution (see Chapter 6.7) — is unlikely to apply or, rather, we don't wish to make that assumption (Chapter 19). The jackknife is a sampling method involving repeatedly sampling from the original data set, but each time leaving one value out. The estimator, for example, the sample mean, is calculated for each sample. The repeated estimates from the jackknife approach yield many estimates which, collected, are used to calculate the sample variance. Jackknife estimators tend to be less biased than those from classical asymptotic statistics.

Bootstrapping, and not jackknife resampling, may now be the preferred permutation approach (e.g., Google Scholar search "bootstrap statistics" 36K hits; "jackknife statistics" 17K hits), but which method is best depends on qualities of the data set. Bootstrapping involves large numbers of permutations of the original data, which, in short, means we repeatedly take many samples of our data and recalculate our statistics on these sets of sampled data. We obtain statistical significance by comparing our result from the original data against how often results from our permutations on the resampled data sets exceed the originally observed results. By permutation methods, the goal is to avoid the assumptions made by large-sample statistical **inference**, i.e., reaching conclusions about the population based on samples from the population. Since its introduction, "bootstrapping" has been shown to be superior in many cases for statistics of error compared to the standard, classical approach (add citations).

There are many advocates for the permutation approaches, and, because we have computers now instead of the hand calculators our statistics ancestors used, permutation methods may be the approach you will take in your own work. However, the classical approach has its strengths — when the conditions, that is, when the assumptions of **asymptotic statistics** are met by the data, then the classical approaches tend to be less **conservative** than the permutation methods. By conservative, statisticians mean that a test performs at the level we expect it to. Thus, if the assumptions of classical statistics are met they return the correct answer more often than do the permutation tests.

Error and the observer

Individual researchers make observations, therefore, we can talk about observer variation as a kind of error measurement. For repeated measures of the same object by an individual, we would expect the individual to return the same results. To the extent repeated measures differ, this is **intraobserver error**. In contrast, measures of the same object from different individuals is **interobserver error**. For a new instrument or measurement system, one would need to establish the reliability of the measure: confronted with the same object, do researchers get the same measurement? Accounting for interobserver error applies in many fields, e.g., histopathology of putative carcinoma slides (Franc et al 2003), liver biopsies for cirrhosis (Rousselet et al 2005), blood cell counts (Bacus 1973).

Confidence in estimates

A really useful concept in statistics is the idea that you can assign how confident you are to an estimate. This is another way to speak of the accuracy of an estimate. Clearly, we have more confidence in a sample estimate for a population parameter if many observations are made. Another factor in our ability to estimate is the magnitude of observation differences. In general, the larger the differences among values from repeated trials, the less confident we will be in out estimate, unless, again, we make our estimates from a large collection of observations. These two quantities, **sample size** and **variability,** along with our level of confidence, e.g., 95%, are incorporated into a statistic called the **confidence interval**.

We will use this concept a lot throughout the course; for now, a simple but **approximate confidence interval** is to use the 2 x **SEM** rule (as long as sample size large): twice the standard error of the mean. Take your estimate of the mean, then add (upper limit) or subtract (lower limit) twice the value of the standard error of the mean (if you recall, that's the standard deviation divided by the square-root of the sample size).

$$\mu = \bar{X} \pm 2 \cdot s_{barX}$$

Example. Consider five magnetic darts thrown at a dart board (28 cm diameter, height of 1.68m from the floor) from a distance of 3.15 meters.

Figure 3.5.1: Magnetic dart board with 5 darts.

The distance in centimeters (cm) between where each of the five darts landed on the board compared to the bullseye is reported in Table 3.5.1.

Table 3.5.1. Results of five darts thrown at a target

Dart label	Distance in centimeters from center
1	7.5
2	3.0
3	1.0
4	2.7
5	7.4

Note:

Use of the coordinate plane, and including the angle measurement in addition to distance (the vector) from center, would be a better analysis. In the context of darts, determining accuracy of a thrower is an Aim-Point targeting problem and part of your calculation would be to get MOA (minute of angle). For the record, the angles (degrees) were

1. 124.4
2. -123.7
3. 96.3
4. -84.3
5. -31.5

measured using imageJ. Because there seems to be an R package for just about every data analysis scenario, unsurprisingly, there's an R package called `shotGroups` to analyze shooting data.

How precise was the dart thrower? We'll use the **coefficient of variation** as a measure of precision. Second, how accurate were the throws? Use R to calculate

```
darts = c(7.5, 3.0, 1.0, 2.7, 7.4)
#use the coefficient of variation to describe precision of the throws
coefVar = 100*(sd(darts)/mean(darts)); coefVar
[1] 68.46141
```

Confidence Interval to describe accuracy

Note that the true value would be a distance of zero — all bullseyes. We need to calculate the standard error of the mean (SEM); then, we calculate the confidence interval around the sample mean.

```
#Calculate the SEM
SEM <- sd(darts)/sqrt(length(darts)); SEM
[1] 1.322649
#now, get the lower and upper limit, subtract from the mean
confidence <- c(mean(darts)-2*SEM, mean(darts), mean(darts)+2*SEM); confidence
[1] 1.674702, 4.320000, 6.965298
```

The mean was 4.3 cm; therefore, to get the lower limit of the interval subtract 2.65 $(2 \cdot SEM = 2.645298)$ from the mean; for the upper limit add 2.65 to the mean. Thus, we report our approximate confidence interval as $(1.7, 7.0)$, and we read this as saying we are about 95% confident the population value is between these two limits. Five is a very small sample number*, so we shouldn't be surprised to learn that our approximate confidence interval would be less than adequate. In statistical terms, we would use the *t*-distribution, and not the normal distribution, to make our confidence interval in cases like this.

*Note:

As a rule, implied by **Central Limit theory** and use of **asymptotic statistical estimation**, a sample size of 30 or more is safer, but probably unrealistic for many experiments. This is sometimes called as the **rule of thirty**. (For example, a 96-well PCR array costs about $500; with $n = 30$, that's $15,000 US Dollars for one group!). So, what about this rule? This type of thinking should be avoided as "a relic of the pre-computer era," (Hesterberg, T. (2008). It's Time To Retire the" n>= 30" rule.). We can improve on asymptotic statistics by applying bootstrap principles (Chapter 19).

We made a quick calculation of the confidence interval; we can get make this calculation by hand by incorporating the *t* distribution. We need to know the **degrees of freedom**, which in this case is 4 ($n - 1$, where $n = 5$). We look up critical value of *t* at 5% (to get our 95% confidence interval), $t = 2.78$. Subtract for lower limit $t \cdot SEM$ and add for upper limit $t \cdot SEM$ to the sample mean for the 95% confidence interval. We can get help from R, by using the one-sample t-test with a test against the hypothesis that the true mean is equal to zero

```
#make an attach a data frame object
Ddarts <- data.frame(darts)
t.test(darts,mu=0)
One Sample t-test
data: darts
t = 3.2662, df = 4, p-value = 0.0309
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
```

```
 0.6477381 7.9922619
sample estimates:
mean of x
 4.32
```

`t.test` uses the function `qt()`, which provides the quantile function. To recreate the 95% CI without the additional baggage output from the `t.test`, we would simply write

```
#upper limit
mean(darts)+ qt(0.975,df=4)*sd(darts)/sqrt(5)

#lower limit
mean(darts)+ qt(0.025, df = 4, lower.tail=TRUE)*sd(darts)/sqrt(5)
```

where `sd(darts)/sqrt(5)` is the standard error of the mean.

Or, alternatively, download and take advantage of a small package called `Rmisc` (not to be confused with the RcmdrMisc package) and use the function `CI`

```
library(Rmisc)
CI(darts)
     upper     mean     lower
7.9922619 4.3200000 0.6477381
```

The advantage of using the `CI()` command from the package `Rmisc` is pretty clear; I don't have to specify the degrees of freedom or the standard error of the mean. By default, CI reports the 95% confidence interval. we can specify any interval simply by adding to the command. For example,

```
CI(darts, ci=0.90)
```

reports upper and lower limits for the 90% confidence interval.

Significant figures

And finally, we should respect **significant figures**, the number of digits which have meaning. Our data were measured to the nearest tenth of a centimeter, or two significant figures. Therefore, if we report the confidence interval as (0.6477381, 7.9922619), then we imply a **false level of precision**, unless we also report our **random sampling error of measurement.**

R has a number of ways to manage output. One option would be to set number of figures globally with the `options()` function — all values reported by R would hold for the entire session. For example, `options(digits=3)` would report all numbers to three significant figures. Instead, I prefer to use `signif()` function, which allows us to report just the values we wish and does not change reporting behavior for the entire session.

```
signif(CI(darts),2)
upper  mean lower
 8.00  4.30  0.65
```

Note:

The `options()` function allows the R user to set a number of settings for an R session. After gaining familiarity with R, the advanced user recognizes that many settings can be changed to make the session work to report in ways more convenient to the user. If curious, submit `options()` at the R prompt and available settings will be displayed.

The R function `signif()` applies rounding rules. We apply rounding rules when required to report estimates to appropriate levels of precision. Rounding procedures are used to replace a number with a simplified approximation. Wikipedia provides a comprehensive list of rounding rules. Notable rules include

- directed rounding to an integer, e.g., rounding up or down
- rounding to nearest integer, e.g., round half up if the number ends with 5
- randomly rounding to an integer, e.g., stochastic rounding.

With the exception of stochastic rounding, all rounding methods impose biases on the sets of numbers. For example, the round half up method applied for numbers above 5, round down for numbers below 5 will increase the variance of the sample. In R, use `round()` for most of your work. If you need one of the other approaches, for example, to round up, the command is `ceiling()` ; to round down we use `floor()` .

When to round?

No doubt your previous math classes have cautioned you about the problems of **rounding error** and their influence on calculation. So, as a reminder, if reporting calls for rounding, then always round after you've completed your calculations, never during the calculations themselves.

A final note about significant figures and rounding. While the recommendations about reporting statistics are easy to come by (and often very proscriptive, e.g., Table 1, Cole 2015), there are other concerns. **Meta-analysis**, which are done by collecting information from multiple studies, would benefit if more and not fewer numbers are reported, for the very same reason that we don't round during calculations.

Questions

1. Calculate the correct 90% and 99% confidence intervals for the dart data using the t-distribution
   - by hand
   - by one alternative method in R, demonstrated with examples in this page
2. How many significant figures should be used for the volumetric pipettor p1000? The p200? The p20 (data at end of this page)?
3. Another function, `round()` , can be used. Try

```
round(CI(darts),2)
```

1. - and report the results: vary the significant figures from 1 to 10 (`signif()` will take digits up to 22).
   - Note any output differences between `signif()` and `round()` ? Don't forget to take advantage of R help pages (e.g., enter `?round` at the R prompt) and see Wikipedia.
2. Compare rounding by `signif()` and `round()` for the number 0.12345. Can you tell which rounding method the two functions use?
3. Calculate the coefficient of variation (CV) for each of the three volumetric pipettors from the data at end of this page. Rank the CV from smallest to largest. Which pipettor had the smallest CV and would therefore be judged the most precise?
4. Standards distinguish between within run precision and between run precision of a measurement instrument. The data in Table 1 were all recorded within 15 minutes by one operator. What kind pf precision was measured?
5. Calculate the standard error of the means for each of the three pipettors from the data provided at end of this page.
6. Calculate the approximate confidence interval using the 2SE rule and judge which of the three pipettors is the most accurate (narrowest confidence interval)
   - Repeat, but this time apply your preferred R method for obtaining confidence intervals.
   - Compare approximate and R method confidence intervals. How well did the approximate method work?

Data sets

Pipette calibration

Table 3.5.2. Mass (grams) of 100 µL of distilled water dispensed by three volumetric pipettes.

p1000	p200	p100
0.113	0.1	0.101

p1000	p200	p100
0.114	0.1	0.1
0.113	0.1	0.1
0.115	0.099	0.101
0.113	0.1	0.101
0.112	0.1	0.1
0.113	0.1	0.1
0.111	0.1	0.1
0.114	0.101	0.101
0.112	0.1	0.1