List of Definitions, Concepts, and Formulas for Math 125

Definitions and Concepts

Chapter 19

Population (defined):

A population is a class of individuals, about which the investigator wants to generalize.

Sample (defined):

A sample is part of a population.

Parameter (defined):

A parameter is a numerical fact about a population. Ususlly a parameter cannot be determined exactly, but can only be estimated.

Statistic (defined):

A statistic can be computed from a sample, and used to estimate a parameter. A statistic is what the investigator knows. A parameter is what the investigator wants to know.

The Accuracy of the Estimate of a Paramter:

When estimating a parameter, one major issue is accuracy: how close is the estimate going to be?

Some methods for choosing samples are likely to produce accurate estimates. When thinking about a sample survey, ask yourself:

- What is the population? the parameter?
- How was the sample chosen?
- What was the response rate?

The best methods of choosing the sample involve the planned introduction of chance.

If the sample was chosen well, it is likely to be representative of the population.

To minimize bias, an impartial and objective probability method should be used to choose the sample.

Chapter 20

The Sample:

The sample is only part of the population, so the percentage composition of the sample usually differs by some amount from the percentage composition of the whole population.

The Definiton of the Standard Error:

For probability samples, the likely size of the chance error (the amount off) is given by the standard error.

The Need for a Box Model:

To figure the SE, a box model is needed. When the population involves classyfing and counting, or taking percents, there should be only 0's and 1's in the box. Change the box, if necessary.

The Standard Error for a Percentage as the Sample Size Changes:

Multiplying the size of a sample by some factor divides the SE for a percentage not by the whole factor—but by its square root.

The Likely Result of the Draws:

When drawing at random from a box of 0's and 1's, the percentage of 1's among the draws is likely to be around its expected value, give or take its SE.

Relating the Size of the Population to the Accuracy of the Sample:

When the sample is only a small part of the population, the number of individuals in the population has almost no influence on the accuracy of the sample percentage. It is the absolute size of the sample (that is, the number of indivduals in the sample) which matters, not the size relative to the population,

The square root law is exact when the draws are made with replacement. When the draws are made without replacement, the formula gives a good approximation—provide the number of tickets in the box is large relative to the number of draws.

There is a correction factor that gets the exact SE when drawing without replacement. When the number of tickets in the box is large relative to the number of draws, the correction factor is nearly one and may be safely ignored.

Technical note: the SE for the percentage may be calculated using an alternate method.

First find the SE for the number: that is the SE for the sum of the draws from the 0–1 counting box.

Then apply the formula: SE for the percentage $= \frac{\text{SE for number}}{\text{size of sample}} \times 100\%$.

Generally, it is enough to use the simpler formula for the SE: in terms of the SD of the box and the number of draws. But if the SE for number has already been found, consider using the alternate one shown above.

The simpler formula is: SE for percentage $= \frac{\text{SD of box}}{\sqrt{\text{number of draws}}} \times 100\%$.

For example: A fair coin is tossed 625 times. Find the standard error for the percentage of heads.

The SD of the 0–1 counting box is $\sqrt{(1/2) \times (1/2)} = 0.5$. The SE for the number of heads equals the square root of the number of tosses times the SD of the box: $\sqrt{625} \times 0.5 = 25 \times 0.5 = 12.5$. The SE for the percentage of heads equals the SE for the number of heads divided by the number of tosses times 100%: $(12.5/625) \times 100\% = 0.2 \times 100\% = 2\%$. This is the alternate method.

Or, by our usual method, the SE for the percentage of heads just equals the SD of the couning box divided by the the square root of the number of draws multiplied by 100%, or $(0.5/\sqrt{625}) \times 100\% = 50\%/25 = 2\%$.

Using the Normal Curve to get Probabilites for Sample Percentages

If a simple random sample is taken from a large population, the normal curve may be used to find chances for the percentage in the sample. The standard units for a particular percentage may be found by subtracting the expected value for the sample percent and then dividing by the standard error for the sample percent.

$$\text{standard units for sample percent} = \frac{\text{observed percentage} - \text{EV\%}}{\text{SE\%}}.$$

Chapter 21

Interpreting a Confidence Interval:

The Chances are in the sampling procedure, not in the parameter.

A confidence interval is used when estimating an unknown parameter from sample dats. The interval gives a range for the parameter, and a confidence level that the range covers the true value. Do not make the mistake of thinking that the confidence level could be used to get probabilites for the parameter.

Chapter 23

The Standard Error for a Sample Average as the Sample Size Changes:

When drawing at random with replacement from a box of tickets, multiplying the number of draws by a factor (like 4) divides the SE for the average of the draws by the square root of that factor ($\sqrt{4} = 2$).

Technical note:
When choosing a formula for the average of the draws, generally prefer the formula:

SE for the average = (SD of the box)/$\sqrt{\text{number of draws}}$.

Think of the formula:

SE for the average of the draws = $\dfrac{\text{SE for the sum}}{\text{number of draws}}$

as an alternate formula that is most helpful in the case where the value of the SE for the sum has already been found. (In the Freeman text, this is used as the main formula, but the preferred version is generally much easier to apply.)

For example: The tickets in a box average to 127 and the SD is 3.75. Two hundred and twenty-five tickets are drawn at random without replacement. Find the standard error for the average of the draws.

The preferred formula gives: $3.75/\sqrt{225} = 3.75/15 = 0.25$.

The alternate formula would require first finding the SE for the sum of draws.
SE for the sum = $\sqrt{\text{the number of draws}} \times$ SD of the box = $\sqrt{225} \times 3.75 = 15 \times 3.75 = 56.25$.
Then the SE for the average of the draws = $\frac{56.25}{225} = 0.25$.

Confidence Intervals for a Population Average

Use the bootstrap to estimate the population SD and, from that, the SE for the sample average.

As before, the confidence interval will be "sample average $\pm z$ SEs," with z based on the confidence level. For a 95%-confidence interval z will equal 2.

Chapter 24

A Model for Measurement Error

The Chances are in the measuring procedure, not the thing being measured.

If the data show a trend or pattern over time, a box model does not apply.

The square root law only applies to draws from a box.

The Gauss Model

In the Gauss model, each time a measurement is made, a ticket is drawn at random with replacement from the error box. The number on the ticket is the chance error. It is added to the exact value to give the actual measurement. The average of the error box is equal to 0.

When the Gauss model applies, the SD of a series of repeated measurements can be used to estimate the SD of the error box. This estimate is good when there are enough measurements.

Statistical inference can be justified by putting up an explicit chance model for the data. No box, no inference.

Example of the Gauss Model follows on the next page.

Example:

Laser altimeters can measure elevation to within a few inches, without bias, and with no trend or pattern to the measurements. As part of an experiment, 144 readings were made on the elevation of a mountain peak. These averaged out to 82,379 inches, and their SD was 30 inches.

Decide whether each of the following is true or false.

(a) Each reading off 82,379 inches by 30 inches or so.

(b) The average of all 144 readings is off 82,379 inches by 2.5 inches or so.

(c) A 95%-confidence interval for the elevation of the mountain peak is 82,374 to 82,784 inches.

(d) There is no such thing as a confidence interval for the average of the 144 measurements.

(e) About 95% of the 144 readings were in the range $82{,}379 \pm 5$ inches.

(f) If a 145th reading were taken and it turned out to be 82,394 inches, that would be quite surprising as the standard error for the average of the readings was only 2.5 inches.

(g) With the Gauss model, the average of the error box is 0, so the expected value of the average of the readings equals the exact height of the mountain peak.

(h) The elevation of the mountain peak is estimated at 82,379 inches; this estimate is likely to be off by 2.5 inches or so.

(i) If the data don't follow the normal curve, you can't use the curve to get confidence levels.

Answers: (a) True (b) False (c) True (d) True (e) False (f) False (g) True (h) True (i) False

Comments:

(a) use SD

(b) The average is 82,379 inches exactly.

(d) the average is known exactly, why estimate it?

(e) That is mixing up SD and SE. Also we do not know whether the readings follow the normal curve; only the average of the 144 draws follows the normal curve.

(f) Not surprising, since the next reading would be off by about 30 inches, the SD.

(g) Each reading equals the exact height plus a draw from the error box.

(i) You use the curve on the probability histogram for the average, not the histogram for the data (see pages 418 and 419).

Formulas

Simple Random Sampling

Simple random sampling means drawing at random without replacement.

Connecting the Statistic from the Sample with the Parameter of the Box

To minimize bias, an impartial and objective probability method should be used to choose the sample.

In general, use the following formula:

$$\text{estimate} = \text{parameter} + \text{bias} + \text{chance error}.$$

The Expected Value for a Sample Percentage

For a simple random sample from a given population, the following equation holds for the percentage (in a specified category)

$$\text{percentage in sample} = \text{percentage in population} + \text{chance error}.$$

With a simple random sample, the expected value for the sample percentage equals the population percentage.

The Standard Error for a Sample Percentage

$$\textbf{SE for percentage} = \left(\textbf{SD of box}/\sqrt{\textbf{number of draws}}\right)\times\textbf{100\%}.$$

It is a fact that the SD of a 0–1 counting box is always .5 or less. This means that the SE for a percentage is always less than $\dfrac{50\%}{\sqrt{\text{number of draws}}}$, no matter what the percentage of the box.

Warning: if the problem involves classifying and counting to get a percent, put 0's and 1's in the box.

Inference from Sample to Population (the Bootstrap Method)

The bootstrap. When sampling from a 0-1 box whose composition is unknown, the SD of the box can be estimated by substituting the fractions of 0's and 1's in the sample for the unknown fractions in the box. The estimate is good when the sample is reasonably large.

Confidence intervals

A confidence interval for a percentage—with a confidence level specified as a percent—is a range of percentages such that you are that percent confident that the population percentage is in that interval.

A confidence interval is based on the results of a single sample of a particular size.

- the interval "sample percentage ± 1 SE" is an approximate 68%-confidence interval for the population percentage.
- the interval "sample percentage ± 2 SEs" is an approximate 95%-confidence interval for the population percentage.
- the interval "sample percentage ± 3 SEs" is an approximate 99.7%-confidence interval for the population percentage.

The Half-Sample Method

Cluster samples are less informative than simple random samples of the same size. So the simple random sample formulas for the standard error do not apply.

The half-sample method may be used to estimate the standard error. Two halves of the sample are used to estimate the parameter. The difference of each estimate from their average is used to estimate the chance error.

The EV and SE for the Average of the Draws

When drawing at random from a box:

$$\text{EV for average of draws} = \text{average of box.}$$

$$\text{SE for average of draws} = \frac{\text{SE for sum}}{\text{number of draws}}.$$

If you wish, you may compute the SE for an average directly from the SD of the box by the formula

$$\text{SE for average} = (\text{SD of box})/\sqrt{\text{number of draws.}}$$

With a simple random sample, the SD of the sample can be used to estimate the SD of the box. This is helpful when calculating confidence intervals for the population average. The estimate is good when the sample is large.

Using Normal Curve to Figure Chances for Average of Draws

When drawing at random from the box, the probability histogram for the average of the draws will follow the normal curve, even if the contents of the box do not. The histogram must be put into standard units, and the number of draws must be reasonably large.

The formula for the standard units is $z = \dfrac{\text{given sample average} - \text{EV for sample average}}{\text{SE for sample average}}$.

DON'T get confused and put SD in the denominator. Units are the same; be careful.

Use of the Bootstrap to Estimate the SD of the Box

With a simple random sample, the SD of the sample can be used to estimate the SD of the box. This is helpful when calculating confidence intervals for the population average. The estimate is good when the sample is large.

Various Standard Errors

For a given box model there are several SEs, each showing the likely size of a certain chance error. The corresponding formulas are:

$$\begin{aligned}
\text{SE for sum} &= \sqrt{\text{number of draws}} \times \text{SD of box} \\
\text{SE for average} &= \frac{\text{SE for sum}}{\text{number of draws}} = \frac{\text{SD of box}}{\sqrt{\text{number of draws}}} \\
\text{SE for count} &= \text{SE for sum, from a 0–1 box} \\
\text{SE for percent} &= \frac{\text{SE for count}}{\text{number of draws}} \times 100\% = \frac{\text{SD of zero-one box}}{\sqrt{\text{number of draws}}} \times 100\%
\end{aligned}$$

The SE for the sum is basic; the other formulas all come from that one. These formulas apply to draws made at random with replacement from a box.

Do not confuse the SD and the SE for the average.

- The SD says how far a number in the box is from average—for a typical number.
- The SE for the average says how far the sample average is from the population average—for a typical sample.

When estimating the average of the box from the average of the sample, the SE shows the likely size of the amount off. It is a give-or-take number.

The Gauss Model for Measurement Error

The Gauss model applies to repeated measurements on some quantity. According to the model, each time a measurement is made, a ticket is drawn at random with replacement from the error box. The number on the ticket is the chance error. It is added to the exact value to give the actual measurement. The average of the error box is equal to 0.

Since the average of the error box is 0 and each measurement is the exact value (a constant) added to the ticket from the error box, the population average weighing equals the exact value.

When two quantities differ by a constant value, they have the same SD. So the SD of the error box is the same as the SD of the measurements. When the Gauss model applies, the SD of a series of repeated measurements can be used to estimate the SD of the error box. The estimate is good when there are enough measurements.

Now, the sample average of the measurements will be off from the population average (the exact value) by the standard error for the sample average. This standard error is found by dividing the estimated SD of the error box by the square root of the sample.