Introduction to Statistics: Formula Sheet*

Luc Hens
Vesalius College

3 November 2016

Scientific notation, resetting the calculator

Calculators and statistical software often use **scientific notation** to display numbers. The letter `E` (or `e`) stands for exponent of 10:

$$\texttt{2.3E+04} = 2.3 \times 10^4 = 2.3 \times 10\,000 = 23\,000$$
$$\texttt{2.3E-04} = 2.3 \times 10^{-4} = 2.3 \times \tfrac{1}{10\,000} = 0.00023$$

Sometimes the *TI-84* returns an **error message**. If nothing else works, **reset** the memory: Mem → Reset → Reset RAM. Resetting erases lists you stored.

Descriptive statistics

Histogram (p. 37). The height of a block over a class interval is the density:

$$\text{density} = \frac{\text{percentage}}{\text{length of interval}}$$

Average (p. 59):

$$\text{average of a list of numbers} = \frac{\text{sum of numbers in list}}{\text{how many numbers there are in the list}}$$

Root-mean-square (p. 66): r.m.s. size of a list = $\sqrt{\text{average of (entries}^2)}$
Standard deviation (p. 71): SD = r.m.s. deviation from average
SD$^+$ (p. 490):

$$\text{SD}^+ = \sqrt{\frac{\text{number of measurements}}{\text{number of measurements} - 1}} \times \text{SD}$$

Shortcut formula for SD (p. 298). When a list has only two different numbers ("big" and "small"), the SD equals

$$\left(\begin{matrix}\text{big}\\\text{number}\end{matrix} - \begin{matrix}\text{small}\\\text{number}\end{matrix}\right) \times \sqrt{\begin{matrix}\text{fraction with}\\\text{big number}\end{matrix} \times \begin{matrix}\text{fraction with}\\\text{small number}\end{matrix}}$$

Standard units (p. 79) = how many SDs a value is above (+) or below (−) the average:

$$\text{value in standard units} = \frac{\text{value} - \text{average}}{\text{SD}}$$

Interquartile range (p. 89) = 75^{th} percentile − 25^{th} percentile

[*] Page numbers refer to Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics (4th edition)*. New York: Norton.

Descriptive statistics for one variable on *TI-84*:
Enter data in a list (e.g., L_1): STAT → EDIT
STAT → CALC → 1-Var Stats L_1
($\bar{x}$ = ave, Sx = SD$^+$, σ_x = SD, n = no. of entries, Q1 = 25th percentile, Q3 = 75th percentile, Med = median)

Correlation coefficient (p. 132). Convert each variable to standard units. The average of the products gives the correlation coefficient:

$$r = \text{average of } [(x \text{ in standard units}) \times (y \text{ in standard units})]$$

Regression line (p. 204):

$$\text{slope} = \frac{r \times \text{SD of } y}{\text{SD of } x}$$

$$y\text{-intercept} = (\text{ave of } y) - \text{slope} \times (\text{ave of } x)$$

Correlation and regression on *TI-84*:
Enter data in two lists (e.g., x in L_1, y in L_2): STAT → EDIT
CATALOG → DiagnosticOn → ENTER → ENTER
STAT → CALC → LinReg(ax+b) L_1, L_2 (x list first, y list second)
(a = slope, b = intercept)

Normal curve

Area. The standard normal curve is a bell-shaped curve with an average of 0 and SD of 1. Use the *TI-84* to find the area under the standard normal curve :
area between −2 and 1 (example 4 p. 83) : DISTR → normalcdf(−2,1)
area to the right of 1 (example 5 p. 83): DISTR → normalcdf(1, 10 ∧ 99)
area to the left of 2 (example 6 p. 84): DISTR → normalcdf(−10 ∧ 99,2)
($10 \wedge 99 = 10^{99}$ is a very large number)
Percentile. Use the *TI-84* to find the 95th percentile of the normal curve (example 10 pp. 90–91): DISTR → invNorm(.95)

Student's t curve

Area. Use the *TI-84* to find the area under Student's t curve with 4 degrees of freedom to the right of 2.2 (pp. 491–492):
DISTR → tcdf(2.2, 10 ∧ 99, 4) ($10 \wedge 99 = 10^{99}$ is a very large number)
The order of arguments is: lower boundary, upper boundary, degrees of freedom.

Probability

Complement rule (p. 223). The chance of something equals 100% minus the chance of the opposite thing.
Multipication rule (p. 229). The chance that two things will both happen equals the chance that the first will happen, multiplied by the conditional chance that the second will happen given the first has happened.
Independence (p. 230). Two things are independent if the chances for the second given the first are the same, no matter how the first turns out. Otherwise, the two things are dependent.

Addition rule (p. 241). To find the chance that at least one of two things will happen, check to see if they are mutually exclusive. If they are, add the chances. If the things are not mutually exclusive, do not add the chances: the sum will be too big.
Factorial: $4! = 4 \times 3 \times 2 \times 1$ (special cases: $1! = 1$; $0! = 1$)
Binomial formula (p. 259). The chance that an event occurs exactly k times out of n is given by the binomial formula:

$$\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

n is the number of trials, k is the number of times that the event is to occur, and p is the probability that the event will occur on any particular trial. The following assumptions should hold: (i) the value of n must be fixed in advance; (ii) p must be the same from trial to trial; (iii) the trials must be independent. *TI-84*: DISTR → binompdf(n, p, k).

Statistical inference

Expected value for a sum (p. 289). The expected value for a sum of draws made at random with replacement from a box equals

$$\text{(number of draws)} \times \text{(average of box)}$$

Standard error (SE) for a sum (p. 291, square root law). When drawing at random with replacement from a box of numbered tickets,

$$\text{SE for sum} = \sqrt{\text{number of draws}} \times \text{(SD of box)}$$

Expected value for a percentage (p. 359). With a simple random sample, the expected value for the sample percentage equals the population percentage.

Standard error (SE) for a percentage (p. 360). First get the SE for the corresponding number (sum of draws from the 0–1 box); then convert to percent, relative to the size of the sample:

$$\text{SE for percentage} = \frac{\text{SE for number}}{\text{sample size}} \times 100\%$$

Expected value and standard error for an average (p. 410). When drawing at random with replacement from a box:

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

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

Confidence interval for the population percentage (p. 360) is obtained by going the right number of SEs either way from the sample percentage. The confidence level is read off the normal curve. For instance, a 95%-confidence interval for the population percentage is:

$$\text{sample percentage} \pm 2 \text{ SEs}$$

Should only be used with large samples and simple random samples.
TI-84: STAT → TESTS → 1-PropZInt; x = number of times the event occurs = percentage × n. To get percentages multiply the boundaries of the obtained confidence interval by 100%.

Confidence interval for the population average (p. 437) is obtained by going the right number of SEs either way from the average of the draws. The confidence level is read off the normal curve. For example, a 95%-confidence interval for the population average is:

$$\text{sample average} \pm 2 \text{ SEs}$$

Should only be used with large samples and simple random samples.
TI-84: STAT → TESTS → ZInterval ($\bar{x}$ = ave, σ_x = SD, n = sample size). If the sample is small use TInterval ($\bar{x}$ = ave, s_x = SD$^+$, n = sample size).

A **test statistic** says how many SEs away an observed value is from its expected value, where the expected value is calculated using the null hypothesis:

$$\text{test statistic} = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

z **test** (p. 479). Should only be used with large samples. Test statistic:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

To compute the observed significance level (P-value) use the normal curve.
TI-84: STAT → TESTS → Z-Test (μ_0 = value of ave in null hypothesis (= 50 on p. 477), $\bar{x}$ = sample ave, σ = SD, n = sample size. Select the appropriate alternative hypothesis: average of box is different from 50 ($\neq \mu_0$), less than 50 ($< \mu_0$), more than 50 ($> \mu_0$)).

t **test** (p. 493). When SD of the box is unknown or the number of observations is small, *and* the histogram of the contents of the box does not look too different from the normal curve, use the t test. First estimate the SD of the box using SD$^+$. Then compute the t test statistic:

$$t = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

To compute the observed significance level (P-value) use Student's t curve with

$$\text{degrees of freedom} = \text{sample size} - 1$$

TI-84: STAT → TESTS → T-Test; as in Z-Test, with Sx = SD$^+$ (or find the area under Student's curve using DISTR → tcdf, see above).