A3SR Math Review

Fall 2019

A Note:

The notes in this packet are not meant to be comprehensive in any way, nor should they be used for learning this material for the first time. Instead, the resources provided here are meant to accomplish two main goals: 1) to give you a sense of the skills/topics that professors will expect you to have learned prior to the start of this program and 2) to jog your memory/refresh your working knowledge of these concepts if you need a quick re-cap. For anything that you haven't learned before or have not seen in many years, the notes provided here may not be sufficient. Suggested outside resources are provided in a separate document, including links to additional practice problems. In designing these materials, we have tried to be concise, and have only included topics that are directly related to concepts covered in your first semester courses. We have also tried to find a balance between things you are expected to understand versus things that you will be expected to calculate. In this program, a lot of complex calculations can be done in R (or using other online tools, like Wolfram Alpha). Most of the time, you will not be asked to do this work by hand (at least not more than once). However, you will find it much easier to understand results in R (and potential errors) if you have a sense of what is going on "under the hood". More in-depth familiarity and experience with these topics is always helpful (if you have additional time for review), and we hope that this packet provides guidance on where to focus your attention.

Properties of Logarithms

Relevant Courses:

-Quantitative Methods
-Generalized Linear Models

Notes

Definition

Logarithms are defined such that $log_b(A) = X$ is equivalent to $b^X = A$

Properties

Using properties of exponents and the definition above, we can derive the following:

1. The Product Rule: $log_b(MN) = log_b(M) + log_b(N)$
2. The Quotient Rule: $log_b(\frac{M}{N}) = log_b(M) - log_b(N)$
3. The Power Rule: $log_b(M^p) = plog_b(M)$
4. $log_b(b^X) = X$
5. $b^{log_b(X)} = X$
6. $log_b b = 1$
7. $log_b 1 = 0$

Example 1: Expanding logarithms

$log_e(\frac{2x^3}{y})$
$= log_e(2x^3) - log_e(y)$
$= log_e(2) + log_e(x^3) - log_e(y)$
$= log_e(2) + 3log_e(x) - log_e(y)$

Example 2: Condensing logarithms

$2log_3(x) + log_3(5) - log_3(2)$
$= log_3(x^2) + log_3(5) - log_3(2)$
$= log_3(5x^2) - log_3(2)$
$= log_3(\frac{5x^2}{2})$

Practice Problems

1. Solve the following:
   a. $log_e(e^x)$
   b. $log_{10}(100)$
   c. $log_{10}(\frac{1}{10})$
   d. $log_{10}(0)$
2. Expand the following:
   a. $log_{10}(\frac{5y^3}{x})$
   b. $log_2(\frac{4y^2}{3x})$
   c. $log_e(2x^2y^3)$
3. Condense the following:
   a. $4log_3(x) - 2log_3(y)$
   b. $log_2(x) + 5log_2(y) - log_2(5)$
   c. $log_{10}(5) + log_{10}(2)$

Answers

1. Solve:
   a. x
   b. 2
   c. -1
   d. There is no solution because there is no power of 10 that would equal 0
2. Expand:
   a. $log_{10}(5) + 3log_{10}(y) - log_{10}(x)$
   b. $2 + 2log_2(y) - log_2(3) - log_2(x)$
   c. $log_e(2) + 2log_e(x) + 3log_e(y)$
3. Condense
   a. $log_3(\frac{x^4}{y^2})$
   b. $log_2(\frac{xy^5}{5})$
   c. 1

Matrix Algebra

Relevant Courses:

-Quantitative Methods

Notes

Definitions

An $m \times n$ matrix A has m rows, n columns, and can be written as:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{m1} & a_{m2} & ... & a_{mn} \end{bmatrix}$$

The transpose of an $m \times n$ matrix, A is the $n \times m$ matrix (denoted A^T) such that every element a_{ij} in matrix A is moved to row j and column i. For example, if:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 5 \\ 3 & 4 & 7 \end{bmatrix}$$

then,

$$\mathbf{A}^T = \begin{bmatrix} 1 & 3 \\ 2 & 4 \\ 5 & 7 \end{bmatrix}$$

The $n \times n$ identity matrix I_n is a matrix with 1s on the diagonal and 0s everywhere else:

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 \\ ... & ... & \ddots & ... \\ 0 & 0 & ... & 1 \end{bmatrix}$$

An $n \times n$ matrix is called "diagonal" if all elements not on the diagonal are zeros. For example:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

An $n \times n$ matrix is called "upper triangular" if all elements below the diagonal are zeros. For example:

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 4 & 2 \\ 0 & 0 & 4 \end{bmatrix}$$

An $n \times n$ matrix is called "lower triangular" if all elements above the diagonal are zeros. For example:

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 4 & 0 \\ 0 & 4 & 4 \end{bmatrix}$$

Suppose that an we have 2 $n \times n$ matrices, $\mathbf{A}$ and $\mathbf{B}$, such that $\mathbf{AB} = \mathbf{I}_n$ (note: this also implies $\mathbf{BA} = \mathbf{I}_n$). Then we say that $\mathbf{B}$ is the inverse of $\mathbf{A}$ (and vice versa) and we can write, $\mathbf{B} = \mathbf{A}^{-1}$. A matrix $\mathbf{A}$ has an

inverse if and only if its determinant is not equal to zero. Note that the determinant of a 2×2 matrix can be calculated as follows (it is not important that you are able to calculate the determinant of a higher dimensional matrix by hand):

$$det(\begin{bmatrix} a & b \\ c & d \end{bmatrix}) = ad - bc$$

Operations with matrices

a. Adding matrices ($\mathbf{A} + \mathbf{B} = \mathbf{C}$): If we add 2 $m \times n$ matrices, $\mathbf{A}$ and $\mathbf{B}$, we get another $m \times n$ matrix C such that $c_{ij} = a_{ij} + b_{ij}$

b. Subtracting matrices ($\mathbf{A} - \mathbf{B} = \mathbf{C}$): If we subtract the $m \times n$ matrix $\mathbf{B}$ from the $m \times n$ matrix $\mathbf{A}$, we get another $m \times n$ matrix $\mathbf{C}$ such that $c_{ij} = a_{ij} - b_{ij}$

c. Multiplying a matrix by a scalar ($c\mathbf{A} = \mathbf{B}$): If we multiply the $m \times n$ matrix $\mathbf{A}$ by a scalar, c, then we get another $m \times n$ matrix $\mathbf{B}$ such that $b_{ij} = c * a_{ij}$

d. Matrix multiplication ($\mathbf{AB} = \mathbf{C}$): Note that it is only possible to compute $\mathbf{AB}$ if the number of columns in matrix $\mathbf{A}$ equals the number of rows in matrix $\mathbf{B}$. If this is the case, then when we multiply an $m \times n$ matrix $\mathbf{A}$ by an $n \times k$ matrix $\mathbf{B}$, we get an $m \times k$ matrix, $\mathbf{C}$, such that $c_{ik} = a_{i1}b_{1k} + a_{i2}b_{2k} + \cdots + a_{in}b_{nk}$

Example 1: Matrix Multiplication

$$\begin{bmatrix} 3 & 4 \\ 2 & 5 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 7 & 2 & 1 \\ 3 & 5 & 2 \end{bmatrix} = \begin{bmatrix} (3*7+4*3) & (3*2+4*5) & (3*1+4*2) \\ (2*7+5*3) & (2*2+5*5) & (2*1+5*2) \\ (1*7+2*3) & (1*2+2*5) & (1*1+2*2) \end{bmatrix} = \begin{bmatrix} 33 & 26 & 11 \\ 29 & 29 & 12 \\ 13 & 12 & 5 \end{bmatrix}$$

Properties of matrix operations

1. $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
2. $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$
3. $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
4. $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$
5. If $\mathbf{A}$ is an $m \times n$ matrix, then $\mathbf{I}_m\mathbf{A} = \mathbf{A}$ and $\mathbf{AI}_n = \mathbf{A}$

Note that, in general $\mathbf{AB} \neq \mathbf{BA}$

Writing a system of equations using matrix notation

Note that, if we have a system of equations:
$y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n$
$y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n$
$\cdots$
$y_m = a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n$

We can re-write these equations much more simply as:
$\mathbf{Y} = \mathbf{AX}$ where $\mathbf{Y}$ is a $1 \times m$ matrix, $\mathbf{A}$ is an $m \times n$ matrix, and X is a $1 \times n$ matrix:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{m1} & a_{m2} & ... & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{bmatrix}$$

Practice Problems

1. Solve the following:
   a. $\begin{bmatrix} 2 & 4 & 2 \\ 1 & 3 & 0 \\ 1 & 6 & 2 \end{bmatrix} + \begin{bmatrix} 1 & 5 & 0 \\ -2 & -3 & 0 \\ 1 & 9 & 5 \end{bmatrix}$
   b. $\begin{bmatrix} 2 & 1 \\ -2 & 2 \\ 4 & 2 \end{bmatrix} \begin{bmatrix} 5 & 2 & -1 \\ 3 & 4 & 2 \end{bmatrix}$
   c. Let $\mathbf{A}=\begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 4 & 4 \\ 1 & 2 \\ 7 & 0 \end{bmatrix}$ Calculate $\mathbf{A}^T\mathbf{B}$
   d. Using the same matrices as in part c, calculate $\mathbf{B}^T\mathbf{A}$
2. Show that **A** and **B** are inverses:
   $A = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 2 & 0 \\ 1 & 1 & 1 \end{bmatrix}$
   $B = \begin{bmatrix} -1 & 0.5 & 1 \\ 1 & 0 & -1 \\ 0 & -0.5 & 1 \end{bmatrix}$
3. Suppose that **A** is a 4x3 matrix and **B** is a $3x8$ matrix.
   a. Does **AB** exist? If so, what are the dimensions of **AB**?
   b. Does **BA** exist? If so, what are the dimensions of **BA**?
4. What is the determinant of $\begin{bmatrix} 1 & -2 \\ 4 & 3 \end{bmatrix}$?

Answers

1. Solve:
   a. $\begin{bmatrix} 3 & 9 & 2 \\ -1 & 0 & 0 \\ 2 & 15 & 7 \end{bmatrix}$
   b. $\begin{bmatrix} 13 & 8 & 0 \\ -4 & 4 & 6 \\ 26 & 16 & 0 \end{bmatrix}$
   c. $\begin{bmatrix} 35 & 10 \\ 13 & 18 \end{bmatrix}$
   d. $\begin{bmatrix} 35 & 13 \\ 10 & 18 \end{bmatrix}$
2. Show that **A** and **B** are inverses:
   $\begin{bmatrix} 1 & 2 & 1 \\ 2 & 2 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0.5 & 1 \\ 1 & 0 & -1 \\ 0 & -0.5 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
3. Suppose that **A** is a 4x3 matrix and **B** is a $3x8$ matrix.
   a. **AB** exists and is 4x8
   b. **BA** does not exist
4. The determinant is $(1 * 3) - (-2 * 4) = (3) - (-8) = 11$

Derivatives

Relevant Courses:

-Probability
-Quantitative Methods

Notes

Definition

The derivative of a function $y = f(x)$ with respect to x is defined as a function giving the instantaneous slope of $y = f(x)$ for any value x. Notationally, a derivative can be written in any of the following ways:

$$f'(x) = y' = \frac{df}{dx} = \frac{d}{dx}(f(x)) = \frac{dy}{dx} = \frac{d}{dx}(y)$$

The second derivative of $y = f(x)$ with respect to x is a function giving the rate of change of the instantaneous slope of $f(x)$. Notationally, it can be represented in any of the following ways (note: 3rd, 4th, etc. derivatives are notated in a similar way, with increasing exponents or $'$s):

$$f''(x) = y'' = \frac{d^2f}{dx^2} = \frac{d^2}{dx^2}(f(x)) = \frac{d^2y}{dx^2} = \frac{d^2}{dx^2}(y)$$

Using derivatives to find local minima and maxima of a function

To find all local minima or maxima of a function $y = f(x)$:

1. Take the first derivative of $f(x)$ with respect to x ($f'(x)$).
2. Set this expression equal to zero and solve for x. These values of x are local maxima and minima.
3. Calculate the second derivative of $f(x)$ with respect to x ($f''(x)$)
4. Plug in the values of x calculated in part 2. If the second derivative is positive, this value of x represents a local minimum; if the second derivative is negative, it is a local maximum

Properties

Properties of derivatives:

1. Sum/Difference rule: $(f(x) \pm g(x))' = f'(x) \pm g'(x)$
2. Constant multiple rule: $(cf(x))' = cf'(x)$ where c is a constant
3. Power rule: If $f(x) = x^n$, then $f'(x) = nx^{n-1}$
4. Product rule: $(f(x)g(x))' = f'(x)g(x) + g'(x)f(x)$
5. Quotient rule (given $g(x) \neq 0$): $(\frac{f(x)}{g(x)})' = \frac{f'(x)g(x) - g'(x)f(x)}{g^2(x)}$
6. Chain rule: $(f(g(x)))' = f'(g(x))g'(x)$

Partial derivatives

For a function of more than one variable (i.e., $f(x, y)$), we can take partial derivatives with respect to each variable. The partial derivative of $f(x, y)$ with respect to x is often denoted by either $\frac{\partial f}{\partial x}$ or f_x. The partial

derivative with respect to y would be denoted by $\frac{\partial f}{\partial y}$ or f_y. The partial derivative with respect to x is calculated by treating any non-x variables as constants when applying the above properties.

Practice Problems

1. Find $f'(x)$ for each of the following. Then compute $f''(x)$ for a-d:
   a. $f(x) = 5x^2 + 3x + 1$
   b. $f(x) = \frac{5}{x^3} + 2x^4$
   c. $f(x) = (3x + 1)^5$
   d. $f(x) = 2x^3(x^2 + 1)$
   e. $f(x) = \frac{2x+1}{x^2-5}$
2. Find the partial derivative with respect to x for each of the following:
   a. $f(x, y) = 3xy^2 + 2x$
   b. $f(x, y) = (xy^4 + 2y)^3$
   c. $f(x, y) = 4x^3 + xy + x^2y^2 + 4x + 2$
3. Find all local minima and maxima for the following function (and note whether they are minima or maxima): $f(x) = \frac{2}{3}x^3 - x^2 - 12x$

Solutions

1. $f'(x)$ and $f''(x)$
   a. $f'(x) = 10x + 3$ and $f''(x) = 10$
   b. $f'(x) = \frac{-15}{x^4} + 8x^3$ and $f''(x) = \frac{60}{x^5} + 24x^2$
   c. $f'(x) = 15(3x + 1)^4$ and $f''(x) = 180(3x + 1)^3$
   d. $f'(x) = 10x^4 + 6x^2$ and $f''(x) = 40x^3 + 12x$
   e. $f'(x) = \frac{2(x^2-5)-2x(2x+1)}{(x^2-5)^2}$ and $f''(x) = \frac{8x(x^2+x+5)-2(2x+1)(x^2-5)}{(x^2-5)^3}$
2. $f_x(x, y) =$
   a. $3y^2 + 2$
   b. $3y^4(xy^4 + 2y)^2$
   c. $12x^2 + y + 2xy^2 + 4$
3. Local minima and maxima
   a. $f'(x) = 2x^2 - 2x - 12 = (2x + 4)(x - 3)$, so setting the first derivative equal to 0, we get $0 = (2x + 4)(x - 3)$, with solutions $x = 3$ and $x = -2$. $f''(x) = 4x - 2$. $f''(3) = 4(3) - 2 = 10$ and $f''(-2) = 4(-2) - 2 = -10$. So, $x = 3$ is a local minimum and $x = -2$ is a local maximum. $f(3) = \frac{2}{3}(3)^3 - (3)^2 - 12(3) = 18 - 9 - 36 = -27$ and $f(-2) = \frac{2}{3}(-2)^3 - (-2)^2 - 12(-2) = \frac{-16}{3} - 4 + 24 = \frac{44}{3} \approx 14.67$

Integrals

Relevant Courses:

-Probability
-Quantitative Methods

Notes

Definition

Indefinite integrals

In the previous section, we calculated the derivative of a function, $f(x)$. Finding an indefinite integral (also called an anti-derivative) involves a simple reversal of this process. The indefinite integral, $F(x)$, of a function, $f(x)$, is defined such that $F'(x) = f(x)$ and is written as follows:

$$\int f(x)dx = F(x) + C$$

where C is a constant.

Definite integrals

The definite integral of $f(x)$ from a to b gives the area under the curve of $f(x)$ on the interval between a and b and is calculated/notated as follows:

$$\int_a^b f(x)dx = F(b) - F(a)$$

Basic properties

Note: There are many other properties of integrals, which may be useful once or twice throughout this program. The following are what you will see most commonly, but you should feel comfortable looking up and using additional properties as needed.

1. $\int cf(x)dx = c\int f(x)dx$ where c is a constant
2. $\int f(x) \pm g(x)dx = \int f(x)dx \pm \int g(x)dx$
3. $\int \frac{1}{x}dx = ln|x| + C$
4. $\int x^n dx = \frac{1}{n+1}x^{n+1} + C$

U substitution

Let g be a differentiable function with range on some interval I, and let f be a continuous function on I. Then, by the reverse of the chain rule:

$$\int f(g(x))g'(x)dx = F(g(x)) + C$$

If you note that an integral can be written in this form, U substitution may be useful: Let $u = g(x)$, then $du = g'(x)dx$ and

$$\int f(g(x))g'(x)dx = \int f(u)du = F(u) + C$$

Example:

$\int 2x(x^2+1)^4 dx$
Let $u = x^2 + 1$. Then $\frac{du}{dx} = 2x$, i.e., $du = 2xdx$
Therefore, $\int 2x(x^2+1)^4 dx = \int u^4 du = \frac{u^5}{5} + C = \frac{(x^2+1)^5}{5} + C$

Example (same thing, but with a definite integral):

$\int_1^2 2x(x^2+1)^4 dx$
Let $u = x^2 + 1$. Then $\frac{du}{dx} = 2x$, i.e., $du = 2xdx$
Note that we are calculating the definite integral from $x = 1$ to $x = 2$, which translates to the integral from $u = (1)^2 + 1 = 2$ to $u = (2)^2 + 2 = 5$
Therefore, $\int_1^2 2x(x^2+1)^4 dx = \int_2^5 u^4 du = \frac{u^5}{5}|_2^5 = \frac{5^5}{5} - \frac{2^5}{5} = 618.6$

Practice Problems

1. Evaluate the following indefinite integrals:
   a. $\int 5x^2 dx$
   b. $\int 2x^3 - \frac{3}{x^2} dx$
   c. $\int (2x-2)(x+3) dx$
   d. $\int \frac{2}{x} dx$
   e. $\int \frac{(ln(x))^2}{x} dx$

2. Evaluate the following definite integrals:
   a. $\int_0^1 4x^3 dx$
   b. $\int_{-2}^2 8x^3 + 3x^2 - 5x + 2 dx$
   c. $\int_1^\infty \frac{2}{x^2} dx$
   d. $\int_1^2 \int_0^2 2x^3 + 3y^3x^2 \; dxdy$
   e. $\int_{-1}^1 (1+x)(2x+x^2) dx$

Answers

1. Indefinite integrals:
   a. $\int 5x^2 dx = \frac{5}{3}x^3$
   b. $\int 2x^3 - \frac{3}{x^2} dx = \int 2x^3 - 3x^{-2} dx = \frac{1}{2}x^4 + 3x^{-1} + C$
   c. $\int (2x-2)(x+3) dx = \int 2x^2 + 4x - 6 dx = \frac{2}{3}x^3 + 2x^2 - 6x + C$
   d. $\int \frac{2}{x} dx = 2\int \frac{1}{x} dx = 2ln|x| + C$
   e. $\int \frac{(ln(x))^2}{x} dx = \frac{(ln(x))^3}{3} + C$ (hint: let u=lnx)

2. Definite integrals:
   a. $\int_0^1 4x^3 dx = x^4|_0^1 = 1^4 - 0^4 = 1$
   b. $\int_{-2}^2 8x^3 + 3x^2 - 5x + 2 dx = (2x^4 + x^3 - \frac{5}{2}x^2 + 2x)|_{-2}^2 = (32 + 8 - 10 + 4) - (32 - 8 - 10 - 4) = 24$
   c. $\int_1^\infty \frac{2}{x^2} dx = \int_1^\infty 2x^{-2} dx = -2x^{-1}|_1^\infty = 0 - -2 = 2$
   d. $\int_1^2 \int_0^2 2x^3 + 3y^3x^2 \; dxdy = \int_1^2 (\frac{1}{2}x^4 + y^3x^3)|_0^2 dy = \int_1^2 8 + 8y^3 dy = (8y + 2y^4)|_1^2 = (16+32) - (8+2) = 38$
   e. $\int_{-1}^1 (1+x)(2x+x^2) dx = \frac{1}{12}u^6|_{-1}^3 \approx 60.67$

Variables: Types and Summaries

Relevant Courses:

-Probability
-Quantitative Methods
-Statistical Computing

Notes

Types of variables:

1. Numerical
   a. continuous (example: height)
   b. discrete (example: population)

2. Categorical
   a. nominal (unordered) (example: responses to: "who are you planning to vote for in the upcoming election?")
   b. ordinal (ordered) (example: responses to: "On a scale from 1-10, how happy are you right now?")

Summarizing variables:

Note: without going into too much detail, it is important that summary statistics be chosen and evaluated with the variable type in mind. For example, if I calculate the mean of a variable which is categorical (nominal) with responses coded as 0s and 1s, the mean will give me the percentage of 1s in my dataset. However, this is no longer the case if the same categorical variable is coded with 1s and 2s.

Summary statistics for central tendency

1. Mean: For a variable x and sample size n, the mean is given by $\frac{\Sigma_{i=1}^{n} x_i}{n}$

2. Median: For a variable x, the median is defined as the central value at which 50% of x-values are smaller and 50% are larger

3. Mode: For a variable x, the mode is the most common value of x. A variable can have multiple modes

Summary statistics for spread

1. Range: For a variable x, the range is given by max(x)-min(x)

2. Variance: For a variable x and population size n, the (population) variance is given by $\sigma^2 = \frac{\Sigma_{i=1}^{n}(x_i-\mu)^2}{n}$ where μ is the mean (i.e., average squared difference between each value of x and the mean).

3. Standard deviation: $\sigma = \sqrt{\sigma^2}$

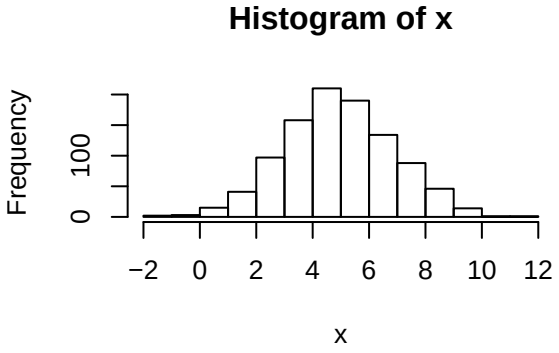
A note about sample variance vs. population variance

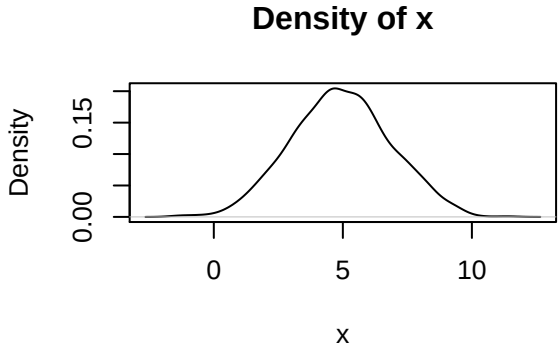
Note that: The definition above is used to calculate "population" variance. When using a sample to estimate the variance of x in a larger population, sample variance is calculated as: $\sigma^2 = \frac{\Sigma_{i=1}^{n}(x_i-\mu)^2}{n-1}$.

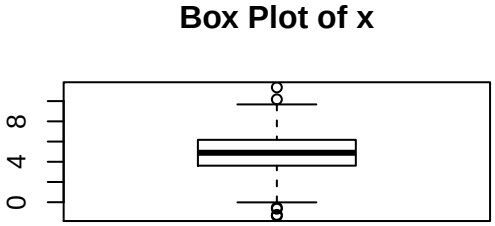
Why? The first formula (where we divide by n) is a "biased" estimator of population variance. The sample variance formula above is not (this is covered in frequentist inference, but is worth a Google search if interested).

Summarizing variables visually

A few common ways of visualizing data (including central tendency, spread, and other characteristics like bimodality and skewness) are demonstrated below (histograms, smoothed density, and box plots). These are worth looking up if they are not immediately familiar.







Basic Probability

Relevant Courses:

-Probability

Notes

Set Theory

This section goes over some definitions and theory before we go any further.

Set - An unordered collection of distinct elements.

A set is made of of elements, and each of those elements is a **member** of the set. We use the $\in$ symbol to denote membership. The following means *element x is a member of set A*.

$$x \in A$$

Curly braces are usually used to denote sets.

$$\{1, 2, 3\}$$

You could think of this set as variable x, given that x is either 1,2, or 3. We write this mathematically as the following (the pipe | means "given")

$$\{x \mid x = 1 \text{ or } x = 2 \text{ or } x = 3\}$$

A more general form of this for any set can be written as: $\{x|conditions\}$ Where x are the elements of the set, defined by whatever conditions the situation calls for.

Some properties of sets

- **A set is uniquely defined by its members** Two sets are equal only if all of their elements are the same. However sets are also unordered - so $\{1, 2, 3\}$ is equal to $\{3, 1, 2\}$
- The **Empty Set** has no contents. It is represented by $\emptyset$ or $\{\}$
- Set A can be a **Subset** of set B if all the elements in set A are also members of set B. Mathematically, this is written as $A \subset B$. Note that the empty set is the subset of any other set and $A \subset A$ for any set A.
- We say that sets $A_1, A_2, A_3, \ldots$ are **mutually exclusive** or **disjoint** if $A_i \cap A_j = \emptyset$ for any distinct pair $A_i \neq A_j$. For instance, in the coin-toss experiment the set A = {Heads} and B = {Tails} would be mutually exclusive.

Sample spaces

For an experiment E, the set of all possible outcomes (sample points) of E is called the **sample space** and is denoted by the letter S

Here are some examples of experiments, sample points, and sample spaces:

a) Experiment: Roll a standard 6-sided die and record the result (the number that is face up)
example of a sample point: 2
sample space: {1,2,3,4,5,6}

b) Experiment: Roll a die, then flip a coin (with probability .6 of landing on heads) and record the die roll and whether heads or tails is facing upward on the coin.
example of a sample point: (2,Heads)
sample space: {(1,Heads),(2,Heads),(3,Heads),(4,Heads),(5,Heads),(6,Heads),(1,Tails),(2,Tails),(3,Tails),(4,Tails),(5,Tails),(6,Tails)}

Events

An event is a specific collection of outcomes, or in other words, a **subset of the sample space**. After the performance of a random experiment, we say that the event A *occurred* if the experiment's outcome *belongs to* A.

Probability Definition

Suppose that a random experiment (also called a random process) is repeated infinitely many times; Then the probability of a particular event is the proportion of times that the event occured.

Law of Large Numbers

As a random experiment is repeated more times, the observed proportion of times that an event occurs will converge to the true empirical probability of that event.

Union, Intersection and Difference

Given subsets A and B, it is often useful to manipulate them in an algebraic fashion. We have three set operations at our disposal to accomplish this: **union**, **intersection**, and **difference**. Additionally we can take the **complement** of one of the sets. Below is a table that summarizes the pertinent information about these operations.

Name	Notation	Definition
Union	$A \cup B$	In either A or B or both
Intersection	$A \cap B$	In both A and B
Difference	$A \setminus B$	In A but not in B

Addition Rule

$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \text{ and } B)$
Note that if A and B are disjoint then $P(A \text{ and } B) = 0$

Multiplication Rule

Generally:
$P(A \text{ and } B) = P(A) \times P(B|A)$
Therefore if A and B are independent:
$P(A \text{ and } B) = P(A) \times P(B)$ because $P(B|A) = P(B)$ if A and B are independent

Complements

For random experiment sample space S: The complement of an event A is denoted A^c and is made up of all sample points in S that are *not* in A.
$A + A^c = 1$

Conditional Probability

The conditional probability of outcome A given outcome B is denoted $P(A|B)$ and is calculated as: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Practice Problems

1. Suppose that you run an experiment that involves flipping 2 fair coins. Consider the following events:
   A=The first coin flip is tails

B=Both coin flips are tails
C=The coin flips are the same (i.e., both heads or both tails)

a. Write out the sample space for this experiment

b. Write out the sample points that comprise each event

c. What is the probability of A and C: $P(A \cap C)$?

d. What is the probability of A or C: $P(A \cup C)$?

e. What is the probability of B given A: $P(B|A)$?

f. Write out the sample points in the complement of B: B^c

2. Suppose that you are playing the following game with your friend: you have a urn with 10 marbles inside. 8 marbles are red and 2 marbles are blue. Your friend randomly picks a marble from the urn. If the marble is blue, your friend wins the game. If the marble is red, then you get to pick a marble from the urn. The game continues until someone draws a blue marble to win the game.

   a. What is the probability of your friend winning on the first turn?

   b. What is your probability of winning on your first turn (i.e., given that your friend got a red marble on the first turn)?

   c. What is the probability that you and your friend both pick red marbles on your first turns?

3. Suppose that you collected responses from a random sample of 200 New Yorkers for the following 2 questions: 1) Do you exercise at least 3 times per week? and 2) Have you ever been diagnosed with Depression? Use the following table (which shows the number of yes/no responses to each question) to calculate the indicated probabilities.

	YesE	NoE
NoD	45	52
YesD	65	38

   a. What is the probability of an individual respondent having been diagnosed with depression?

   b. Based on this sample, what is the probability of an individual respondent having been diagnosed with depression given that they exercise at least 3 times per day?

   c. Based on this sample, what is the probability of an individual respondent having been diagnosed with depression given that they DO NOT exercise at least 3 times per day?

   d. Based on your responses above, do you think that there is an association between exercise frequency and probability of being diagnosed with depression?

Answers

1. Question 1
   a. Sample space = {TT, TH, HT, HH}
   b. A={TT, TH}, B={TT}, C={TT, HH}
   c. $A \cap C = \{\text{TT}\}$ so $P(A \cap C) = \frac{1}{4}$
   d. $A \cup C = \{\text{TT, TH, HH}\}$ so $P(A \cup C) = \frac{3}{4}$
   e. $P(B|A) = \frac{1}{2}$

f. B = {TH, HT, HH}

2. Question 2
   a. $\frac{2}{10}$
   b. $\frac{2}{9}$
   c. $\frac{8}{10} * \frac{7}{9} \approx 0.62$

3. Question 3
   a. $\frac{65+38}{200} = .515$
   b. $\frac{38}{38+52} \approx .42$
   c. $\frac{65}{65+45} \approx .59$
   d. Yes, there seems to be a higher probability of depression among individuals who do not exercise. Note that we could run a chi square test for independence here to test the hypothesis that answers to these questions are independent. This is worth looking up if you have never seen it before.

Random Variables and Probability Density/Mass Functions

Relevant Courses:

-Quantitative Methods
-Probability
-Statistical Computing -Causal Inference

Notes

Random variable definition

Random variables are functions with *numerical* outcomes that occur with some level of uncertainty. Often, experiments with non-numerical outcomes are represented by random variables. For example, flipping a fair coin one time could be considered a random variable with two possible outcomes: 0 (corresponding to tails) or 1 (corresponding to heads).

Random variables can be discrete or continuous:

1. Discrete random variables have a finite number of potential outcomes

2. Continuous random variables have infinitely (countably or uncountably) many possible outcomes

Probability Distributions

A probability distribution of a random variable determines the relative likelihood of each potential outcome of that random variable.

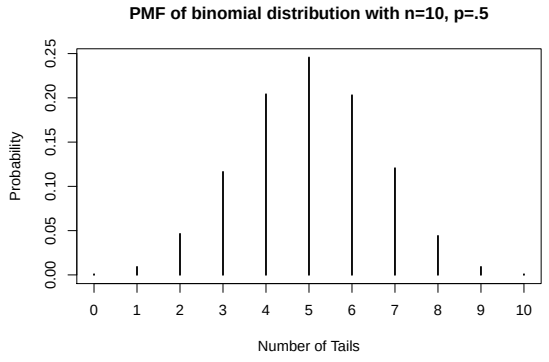
1. Probability Mass Function (PMF): (for discrete random variables) maps possible values of a random variable to their probabilities.

2. Probability Density Function (PDF): (for continuous random variables) area under the curve between any two values gives the probability of the random variable's outcome being between those two values.

There are many common probability distributions that you will need to know (these are mostly covered in Probability), but two particularly common ones are given as examples below. (Note: the major distributions covered in Probability are Bernoulli, Binomial, Normal, Uniform, Geometric, Negative Binomial, Hypergeometric, Poisson, Exponential, Gamma, Beta, Chi-Square, and Student-t. These are important to review if you are not taking Probability, and are worth looking up if you have never seen them before).

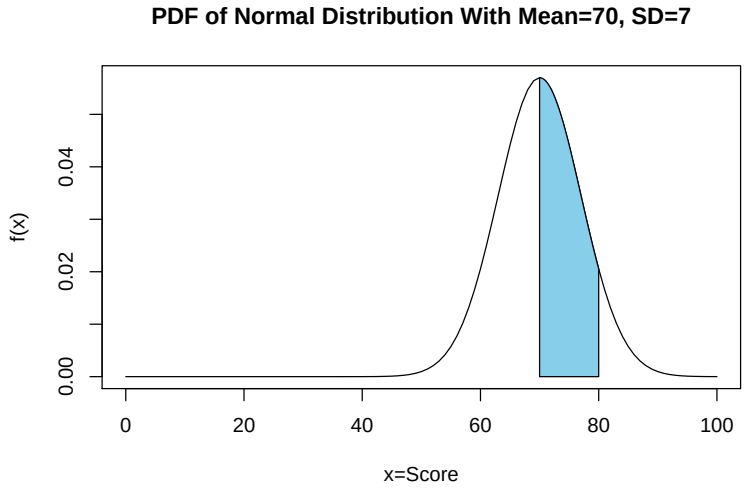
PMF Example

Binomial(n=10,p=.5) distribution: Can be thought of as the probability of x tails among 10 flips of a fair coin



PDF Example

Another common probability distribution is the normal distribution. Suppose that you are told that scores on a particular test are (approximately) normally distributed with mean 70 and standard deviation 7 (for the purposes of this thought experiment, suppose that test scores are continuous, not discrete). Then, the area of the blue shaded area in the distribution below would give the approximate probability of any particular student earning between a 70% and 80%.



Note:
It is helpful to remember the following for any data that follows a normal distribution:
-Approximately 68% of the data falls within 1 standard deviation of the mean
-Approximately 95.5% of the data falls within 2 standard deviations of the mean
-More than 99% of the data falls within 3 standard deviations of the mean

Z-Scores

Often, we want to compare two numbers and understand the extent to which they are different. One way to do this is by using Z-scores. Z-scores are generally used for observations that follow an approximately normal distribution. The Z-score for a particular observation x from a distribution with mean μ and standard deviation σ is given by $Z = \frac{x-\mu}{\sigma}$. The Z-score for any observation tells you how far (in standard deviations) that observation is from the mean, μ. If the observation came from a normal distribution, then Z-scores follow a standard ($\mu = 0, \sigma = 1$) normal distribution and a Z-table can be used to convert the Z-Score to a percentile (or vice versa). You will learn to use a Z-table in probability, but you might want to look this up if you have never seen it before. "Standardizing" a dataset involves converting all values to their Z-Scores.

Example

Suppose test scores on a particular exam are normally distributed with mean 70 and standard deviation 10. Mike gets a score of 60 and Sarah gets a score of 90. Then Mike's Z-score would be $\frac{60-70}{10} = -1$, indicating that he scored 1 standard deviation below average. Using a Z-table, we could determine that this is about 16th percentile. Meanwhile, Sarah's Z-score is $\frac{90-70}{10} = 2$, indicating that she scored 2 standard deviations above average, which is approximately the 98th percentile.

Expectation and Variance

Random variables have expectation (also called expected value) and variance:
For a continuous random variable, X, on the interval [a,b] and PDF $f(x)$, $E(X) = \int_a^b xf(x)dx$
For a discrete random variable, X, with possible outcomes $x_1, ..., x_n$, $E(X) = \Sigma_{i=1}^n x_i P(x_i)$

The variance of a random variable X is given by: $E(X^2) - (E(X))^2$
Note: For now, the exact formulas for expectation and variance of a random variable are less important than

understanding that expectation is a measure of central tendency and variance is a measure of spread.

Properties of expected value:
1. $E(X+Y) = E(X) + E(Y)$
2. $E(cX) = cE(X)$ where c is a constant

Properties of expected value (for a random variable X):
1. $E(X+Y) = E(X) + E(Y)$ where Y is also a random variable
2. $E(X+c) = E(X) + c$ where c is a constant
3. $E(cX) = cE(X)$ where c is a constant

Properties of variance (for a random variable X)
1. $E(X+Y) = Var(X) + Var(Y)$ where Y is also a random variable *ONLY* if X and Y are independent
2. $Var(X+c) = Var(X)$ where c is a constant
3. $Var(cX) = c^2 Var(X)$ where c is a constant

Central Limit Theorem Introduction

Relevant Courses:

-Quantitative Methods
-Statistical Computing
-Probability -Causal Inference

Notes

Central Limit Theorem

If we take repeated independent samples of size n from a population of interest and use the sample mean (denoted $\bar{x}$) to estimate the population mean (denoted μ), then these sample means will be (approximately, if sample size is large enough) normally distributed with mean μ and standard deviation (which we actually call the *standard error* of $\bar{x}$) $\frac{\sigma}{\sqrt{n}}$ where σ is the population standard deviation (which we can estimate using a sample standard deviation if we don't know it). This is true for any population distribution (uniform, skewed, etc.) as long as sample size is large enough.

Note: in general, the term "standard error" is used to describe the standard deviation of an estimate.

Another note: "large enough" sample size is generally (and somewhat randomly) considered to be $n \geq 30$.

A final note: The distribution of sample means described above is called the "sampling distribution" of the mean.

Confidence Intervals (an example)

Suppose that we want to know the average salary of "Data Scientists". To estimate this average salary, we find a sample of 400 people whose title is "Data Scientist" and we ask them to report their salary. Then, we inspect the resulting salaries and find that the mean is 100,000 dollars and standard deviation is 10,000 dollars. Leveraging what we know (at least from this sample) about the variability in salaries, we may want to make a statement about how confident we are that the mean salary of the larger population is in some range (centered around 100,000 dollars).

Because of the central limit theorem, we can use the sample mean plus or minus (approximately) 2 times the standard error of $\bar{x}$ (remember: $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$) to get a confidence interval around an observed sample mean. Even more accurately, we can use: $\bar{x} \pm 1.96 * SE$ because 95% of the area under the curve of the normal distribution PDF lies within 1.96 standard deviations from the mean. So, in the example above, we could say that we are (approximately) 95% confident that the average salary of data scientists is between $100,000 - 1.96 * \frac{10,000}{\sqrt{400}} = 99,020$ and $100,000 - 1.96 * \frac{10,000}{\sqrt{400}} = 100,980$.

Z-Tests, T-Tests, and P-Values

Relevant Courses:

-Quantitative Methods
-Statistical Computing
-Causal Inference

Notes

Hypothesis testing: framework

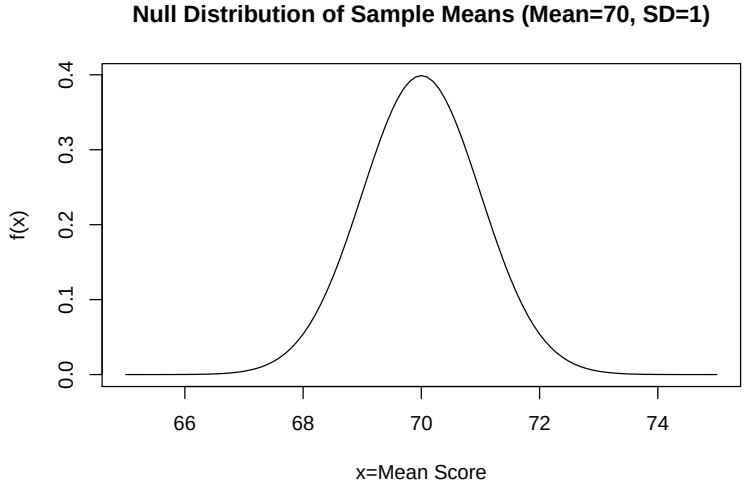
Suppose we know that test scores on a state-wide exam (which is meant to measure math ability) are normally distributed with mean 70 and standard deviation 10. However, one school believes that their students have higher average math ability (as measured by this test). To test this claim, they randomly sample 100 students to take the test, and these students earn a mean score of 72. Is there sufficient evidence to support the school's claim that their students have higher average math ability as measured by this test?

We could design the following null and alternative hypotheses:
Null hypothesis: Students at this school have the same average math ability as measured by the test as the general population.

Alternative hypothesis: Students at this school have higher than average math ability as measured by this test than the general population.

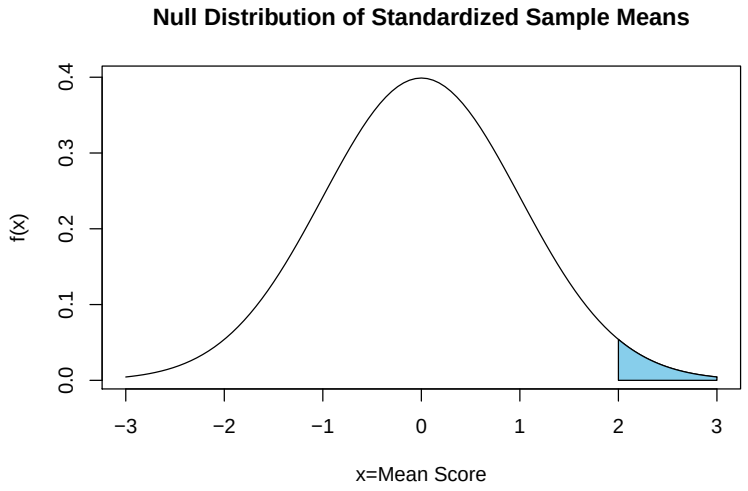
Z-Tests and P-Values

By the central limit theorem: if the null hypothesis is true (i.e., if these students were randomly sampled from the full population), the expected distribution of sample means (with sample size=100) should be normal with mean 70 and standard error $\frac{10}{\sqrt{100}} = 1$. This distribution is shown below:



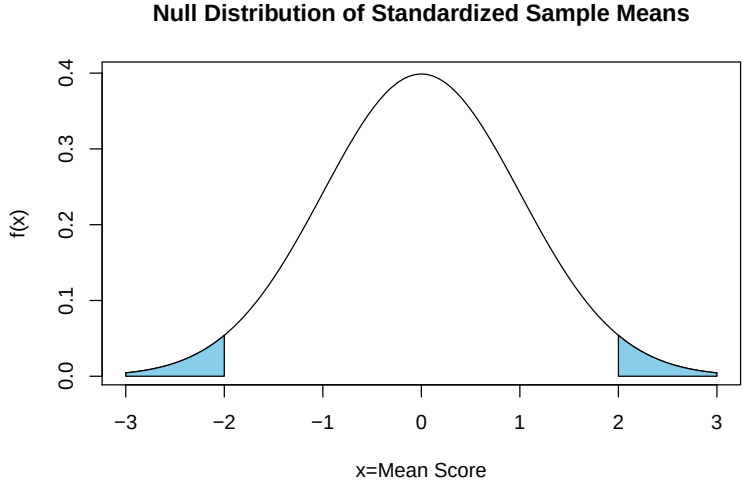
Given this distribution, the probability that we would randomly observe a sample of 100 students with a mean score of 72 can be calculated as follows:

Calculate a test statistic (in this case a Z-score), to determine how many standard deviations above or below the null value (in this case, the population mean) the observed sample mean is: $Z = \frac{72-70}{1} = 2$. Then, we use percentiles for a standard normal distribution to calculate the proportion of samples of size 100 that we would expect to have a sample mean equal to or greater than the observed sample mean by random chance (i.e., area of the blue shaded region below):



In this case, the shaded area is approximately 0.023 (I used 1-pnorm(2) in R to calculate this). Usually, when running a test like this, we would set a significance level ahead of time, and p-values below that value would cause us to "reject the null hypothesis in favor of the alternative". A commonly chosen significance level is (arbitrarily) 0.05. At that level, we would reject the null in this case.

Note:
-The example described above is a "one-tailed" test. A two tailed test in this example would be used to test whether scores at this school are different from the full population (either greater than or less than). A two tailed p-value for this example would equal the shaded area below, which is the proportion of sample means more than 2 points greater or less than the state average that would be expected by random chance under the null. Because the normal distribution is symmetric, this is $0.023 * 2 = .046$:



The t distribution and T-tests

-For sample sizes, $n < 30$, then there are two problems:

1. If the population distribution is skewed, we need a larger sample size to ensure that $\bar{x}$ is normal.

2. If sample size is small, we may not be able to accurately estimate the standard error based on the sample.

If the population distribution is normal, concern 1 goes away. And, to manage concern 2, common practice is to use the t distribution with $n - 1$ degrees freedom to estimate the sampling distribution of $\bar{x}$ (instead of a normal distribution) because the t distribution has heavier tails for small degrees freedom. Note: the

t distribution is a probability distribution with a single parameter, degrees of freedom (df). It is similar to a standard normal distribution, but has heavier tails when df$<$ 30. When df$>$ 30, the t distribution is essentially identical to a standard normal distribution.

One Sample T-tests

A one sample T-test for a sample mean is generally used when sample size is small and observations appear (roughly) normally distributed.

Example:
Based on a national survey, US adults get an average of 6.8 hours of sleep per night. Suppose that we collect a sample of 25 NYU students and find that they got an average of 6.4 hours of sleep per night (sd=1). Is there sufficient evidence to suggest that NYU students are getting less sleep than the national average?

We estimate the standard error of the mean: SE=$\frac{1}{\sqrt{25}} = 0.2$. Then, we construct a test statistic (in this case, a T-score, which is similar to a Z-score) comparing our observed mean with the national average: $T = \frac{6.8-6.4}{0.2} = 2$. Then, we calculate a one-tailed p-value using a t distribution with $n - 1 = 25 - 1 = 24$ degrees of freedom and get 0.028 (I used 1-pt(2,df=24) in R to calculate this). At a significance level of 0.05 we reject the null in favor of the alternative.

T-test for Paired Data

Suppose we want to compare the prices of common coffee drinks at Starbucks and Dunkin' Donuts. For every drink (i.e., small hot coffee, small iced coffee, etc.) we have the price at Starbucks and the price at Dunkin' Donuts. A mock-up of some data (note: this is completely made up) is shown below.

	Starbucks Price	DD Price
Small Hot Coffee	2.75	2.40
Medium Hot Coffee	2.85	2.80
Large Hot Coffee	3.10	3.25
Small Cold Coffee	2.90	2.60
Medium Cold Coffee	3.15	2.90
Large Cold Coffee	3.50	3.75

In this case, we could make a third column in our data, which is calculated by Starbucks Price - DD Price:

	Starbucks Price	DD Price	Price Difference
Small Hot Coffee	2.75	2.40	0.35
Medium Hot Coffee	2.85	2.80	0.05
Large Hot Coffee	3.10	3.25	-0.15
Small Cold Coffee	2.90	2.60	0.30
Medium Cold Coffee	3.15	2.90	0.25
Large Cold Coffee	3.50	3.75	-0.25

Now, we could test the following:
Null hypothesis: The mean price difference is 0
Alternative hypothesis: The mean price difference is not 0

We can use a 2-tailed T-test on the price difference column as follows: Suppose that, for 25 total drinks, the mean price difference was .09 (sd=.25). Then we estimate standard error as: $\frac{.25}{\sqrt{25}} = .05$ and the T-score is $\frac{.09-0}{.05} = 1.8$, giving a p-value of .04 (using 1-pt(1.8,df=24)). So, at a .05 level, we reject the null that the mean price difference is 0.

Two-Sample T-Test Comparing Sample Means (non-paired data)

If we want to compare two independent sample means for data that is not paired (for example, if we had prices for 100 drinks at Starbucks and 100 drinks at Dunkin Donuts, but they were not paired by drink type), we can use an independent sample T-test.

This is done in much the same way as before, except now we need to calculate $\bar{x_1}$ (the mean of group 1), $\bar{x_2}$ (the mean of group 2), calculate their difference, and then construct a T-test comparing this value to a null hypothesis that $\bar{x_1} - \bar{x_2} = 0$. Also, we need a different way of estimating the standard error and need to know how many degrees of freedom to use for our t distribution (see below).

Standard error:
Using what we know about the properties of variance, we can derive that:

$$SE^2(\bar{x_1} - \bar{x_2}) = SE^2(\bar{x_1}) + (-1)^2 SE^2(\bar{x_2}) = SE^2(\bar{x_1}) + SE^2(\bar{x_2}) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

So, $SE(\bar{x_1} - \bar{x_2})$ can be estimated using:

$$SE(\bar{x_1} - \bar{x_2}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Alternatively, if we think that the two groups have equal variances, we can used the pooled standard deviation formula to get a better estimate of $SE(\bar{x_1} - \bar{x_2})$. The formula is:

$$s_{pooled}^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

Degrees of freedom:
In general, we use the smaller of $n_1 - 1$ and $n_2 - 1$ as the degrees of freedom for the t distribution (or if assuming equal variances and using the s_{pooled}^2, we can use $n_1 + n_2 - 2$ degrees of freedom).

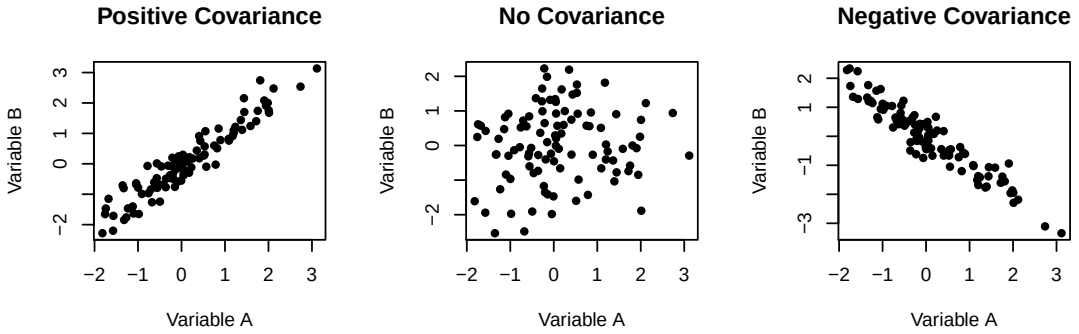
Correlation and Covariance

Relevant Courses:

-Quantitative Methods
-Probability

Notes

Correlation and covariance both measure the extent to which a change in one variable is related to a change in another variable. For example, height and weight tend to have positive covariance and correlation. People who are taller also tend to weigh more (although not always). Conversly, temperature and heating costs tend to have negative covariance and correlation. As the outside temperature increases, the cost to heat my apartment tends to decrease.



Covariance ranges from negative infinity to positive infinity and is calculated as follows (note: when using sample covariance to estimate population covariance, we divide by $n-1$ instead:

$$Cov(x,y) = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Correlation is a standardized version of covariance, which ranges between -1 to 1. A correlation of 0 means the variables are not associated. A correlation of 1 or -1 indicates that the variables are perfectly positively or negatively correlated (i.e., perfectly linearly related). Correlation is calculated as follows:

$$Corr(x,y) = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

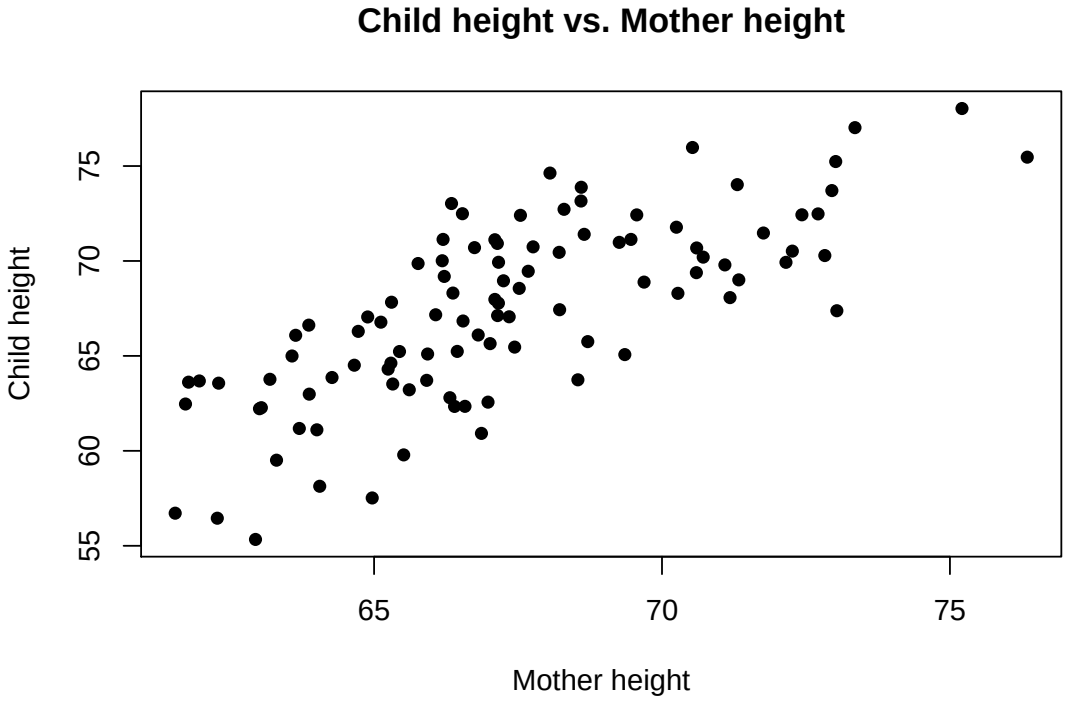
Simple Ordinary Least Squares Regression

Relevant Courses:

-Quantitative Methods

Notes

Suppose that we have the following information for 100 US adults: 1) their height in inches and 2) their mother's height in inches. A plot of the data looks like this:



We see that there appears to be a positive, linear association between mother's heights and children's heights. In other words, on average, taller mothers appear to have taller children. There are many ways we could try to fit a line to this data, but one of the most common is ordinary least squares regression:

1) We set up the following template for a linear relationship between child height and mother height: $\hat{y} = \hat{\beta_0} + \hat{\beta_1} x$ where $\hat{y}$ represents the "predicted" height of a child born to a mother of height x. The $\hat{}$ symbol is used to denote a prediction/estimate. Also, the true values of β_0 and β_1 (for all mothers and children) are unknown so we are estimating them based on this sample as well

2) Based on this "model", the ith child's height is given by $y_i = \hat{\beta_0} + \hat{\beta_1} x_i + \varepsilon_i$ where x_i is their mother's height and ε_i is the difference between the child's true height and their predicted height based on the linear model in part 1. (in other words: $\varepsilon_i = y_i - (\hat{\beta_0} + \hat{\beta_1} x_i) = y_i - \hat{y_i}$

3) Using calculus, we find values of β_0 and β_1 that minimize $\Sigma_{i=1}^{n} \varepsilon_i^2$ (called the residual sum of squares or RSS)

Graphically, this means that the OLS regression line minimizes the vertical distance between each point and the line: