Math Camp, Part II

Harvard Public Policy, Health Policy, and Business Administration PhD Math Camp

17-19 August 2022

Jacob Bradt[†]

Contents

[†] These notes are a slight adaptation of previous versions by Robert French, Rebecca Sachs, Todd Gerarden, Wonbin Kang, Sam Richardson, and Jonathan Borck. I am deeply indebted to previous Math Camp TF's for their work on these notes.

1 Constrained Optimization (Day 1)

Sources for this section can be found in Chiang and Wainwright, chapter 12; Simon and Blume, chapters 18-19, 21; and Martin Osborne's economic math *website*.

1.1 Overview of Constrained Optimization and Notation

Some core concepts:

Unconstrained Optimization: Find the optimal level of one or more "choice variables" that will either maximize or minimize the "objective function."

Constrained Optimization: Same thing, but we have one or more "constraints" that impose limits on what value(s) our "choice variable(s)" can take.

Examples:

Maximizing a utility function subject to a budget constraint (one equality constraint):

$$\max_{x,y} U(x,y)$$

$$\text{s.t. } p_x x + p_y y = w.$$

Profit maximizing for a competitive firm (multiple inequality constraints):

$$\max_{x,y} \Pi(x,y)$$

$$\text{s.t. } \Pi(x,y) \geq 0, x \geq 0, y \geq 0$$

Notation:

$$\max_{x,y} f(x,y)$$

$$\text{s.t. } g(x,y) = c$$

What is the objective function and what is the constraint?

General approach: We work with differentiable functions and use techniques of calculus to solve for optima. This handout emphasizes the use of first order conditions to identify interior optima for optimization problems with equality constraints. If time permits, we will discuss second order conditions, boundary solutions, and inequality constraints.

1.2 Method 1: The Substitution Method

The general approach here is to convert a constrained optimization problem into an unconstrained optimization problem:

1. Use the constraint to solve for one variable in terms of the other(s).
2. Substitute the expression from Step 1 into the objective function.
3. Solve this new unconstrained optimization problem as before:
   - Take the first-order condition(s) to find the potential maxima or minima;
   - Check the second-order condition(s) to verify what each candidate solution is; and
   - Take the $\arg\max(\min)$ of the unconstrained function as the $\arg\max(\min)$ of the constrained function.

Example:

$$\max_{x,y} U(x,y) = 2x + 5\ln y$$
$$\text{s.t. } 6x + 3y = 51$$

Example:

$$\min_{a,b} C(a,b) = (3a-7)^2 + 4b$$
$$\text{s.t. } -24a - 8b = -42$$

1.3 Method 2: The Lagrangian Method

This approach incorporates the constraint into the objective function by introducing a new variable, the Lagrange multiplier λ.

Consider the following setup: We have an objective function $z = f(x, y)$ subject to the constraint $g(x, y) = c$, where $c \in \mathbb{R}$ is a constant. To maximize $f()$ subject to the constraint, we have the following steps:

1. Introduce the Lagrange multiplier, λ, and rewrite the constraint with everything on one side of the equation: $g(x, y) - c = 0$.
2. Create the Lagrangian function, a modified version of the objective function:[1]
$$\mathcal{L} = f(x, y) - \lambda[g(x, y) - c]$$
3. Solve this unconstrained optimization problem as usual, treating the Lagrange multiplier, λ, as an additional variable.
4. Check your solution from Step 3 to determine if it's a maximum or minimum.[2]

Example:

$$\max_{x,y} U(x, y) = 2x + 5 \ln y$$
$$\text{s.t. } 6x + 3y = 51$$

Example:

$$\max_{x,y} f(x, y) = xy$$
$$\text{s.t. } x + 4y = 16$$

[1] Note that some texts use an alternative sign convention: $\mathcal{L} = f(x, y) + \lambda[g(x, y) - c]$

[2] See Simon and Blume, chpater 19.

1.4 Interpreting the Lagrange Multiplier

The Lagrange multiplier is often called the "shadow value" or "shadow price" of the constraint.

It expresses how much the objective function changes if we "relax" the constraint a little bit. Or, a measure of the sensitivity of the optimal value of the objective function to changes in the constraint.

Economic Interpretation:

Utility maximization: The Lagrange multiplier (*when on a budget constraint*) is interpreted as the shadow price of wealth or the marginal utility of wealth - the change in utility that would result from an infinitesimal increase in wealth.

Profit maximization: The Lagrange multiplier (*when on the cost function*) for a particular input is interpreted as the shadow price of that input - the change in profits that would result from an infinitesimal increase in use of that good.

Derivation: Consider the optimization problem, $\max_{x,y} f(x,y)$ s.t. $h(x,y) = a$. Show that,

$$\lambda^*(a) = \frac{d}{da} f(x^*(a), y^*(a)),$$

where x^* and y^* denote the values of x and y that maximize the objective function subject to the constraint.[3]

Another question: What does it mean if $\lambda^* = 0$?

[3] Hint: it may be useful to define the function $F(a) = h(x^\star(a), y^\star(a)) - a$, and note that $\frac{\partial F}{\partial a}$ will always equal zero.

1.5 Convex and Non-Convex Sets

A ***set*** is a collection of objects (often called elements). These objects may indeed be numbers.

Examples:

In one-dimensional Euclidean space, a line segment or series of line segments: $(0,1); \{0,1\}; \{(0,1), 1, [1,3)\}$. Note here how intervals are either defined as a set or an element in a set.

This notion of sets applies to higher dimensional Euclidean space. Moreover, sets do not have to contain elements in Euclidean space; here is a set with three objects, for example: {red, white, blue}.

One (of many) ways in which sets and set notation often comes up is in the context of **level sets**. You will use level sets to study two fundamental functions of microeconomics: production and utility functions. Level sets provide an intuitive way of understanding a function that maps from $\mathbb{R}^n \to \mathbb{R}^1$, describing all combinations of n inputs that produce a given function value. For example, consider the simple Cobb-Douglas production function $Q = f(x,y) = x \cdot y$ where x and y measure amounts of two inputs and Q is output. We can think of the various combinations of x and y producing a fixed level of output Q as depicted by the two-dimensional curve in the $x - y$ plane as shown in the figure below.

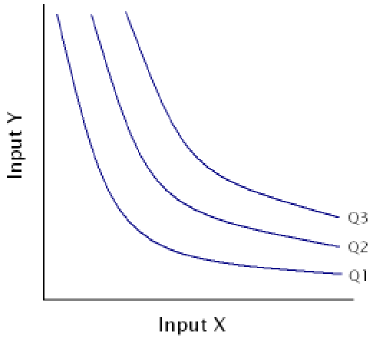


Figure 1: Level sets describing the combinations of inputs x and y that produce a given level of output (Q_1, Q_2, Q_3). The level sets of a production function are also called "isoquants".

A ***convex set*** in Euclidean space is a set $\in \mathbb{R}^n$ where the line segment joining any two points in the set is contained entirely within the set.[4] In other words,

[4] $\mathbb{R}^n$ denotes the n dimensional Euclidean space.

any weighted combination of two points in the set is also in the set.

Algebraically, a set, call it C, is convex if and only if $\forall t \in [0,1]$, and $\forall x, y \in C$, we have that $tx + (1-t)y \in C$.

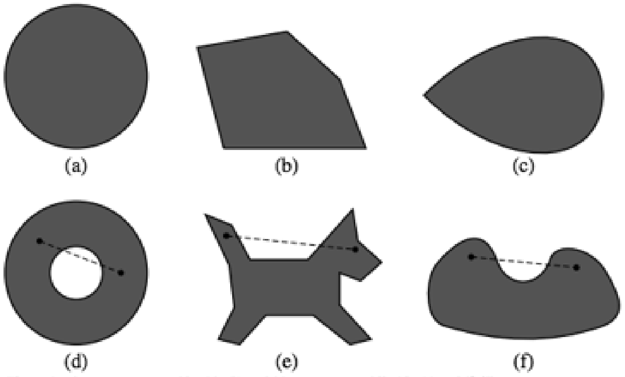


Figure 2: a, b, and c are convex sets; d, e, and f are non-convex.

1.6 Concave and Convex Functions

There are special classes of functions called concave and convex functions. We have the following definitions of each:

- **Concave function:** A real-valued function f defined on a convex subset $U \subset \mathbb{R}^n$ is concave if $\forall x, y \in U$ and $t \in [0,1]$

$$f(tx + (1-t)y) \geq tf(x) + (1-t)f(y)$$

- **Convex function:** A real-valued function g defined on a convex subset $U \subset \mathbb{R}^n$ is convex if $\forall x, y \in U$ and $t \in [0,1]$

$$g(tx + (1-t)y) \leq tg(x) + (1-t)g(y)$$

Why do we care about whether functions are concave or convex? Concavity and convexity imply several desirable properties:

- Let f be concave (convex) function defined on $U \subset \mathbb{R}^n$. If x^* is a critical point[5] of f then $x^* \in U$ is a global maximizer (minimizer) of f on U.
- Let $f_1, ..., f_k$ be concave (convex) functions each defined on the same subset $U \subset \mathbb{R}^n$ and let $a_1, ..., a_k > 0$. Then $a_1 f_1 + ... + a_k f_k$ is a concave (convex) function on U.
- Let f be a function defined on a convex set $U \subset \mathbb{R}^n$. if f is concave then for every $x_0 \in U$, the set

$$C^+_{x_0} \equiv \{x \in U : f(x) \geq f(x_0)\}$$

is a convex set

Concave functions are common in economics. For example, the expenditure and cost functions are concave:

- Expenditure function:

$$e(p, u) = \min\{p_1 x_1 + ... + p_n x_n : u(x) \geq u\}$$

- Cost function:

$$c(w, y) = \min\{w_1 x_1 + ... + w_n x_n : g(x) = y\}$$

The properties of concave (convex) functions are very useful; however, concave functions have a clear downside in economic analysis: concavity is a **cardinal** property, whereas many concepts—importantly, utility—are ordinal.

[5] Meaning $\frac{\partial f}{\partial x_i}(x^*) = 0, \forall i = 1, ...n$

→ Concavity depends on the numbers which the function assigns to the level sets, not just on the shape of the level sets

→ In other words, a monotonic transformation of a concave function need not be concave. But what is a monotonic transformation?

1.7 Positive Monotonic Transformations

We typically apply monotonic transformations to convert difficult-to-analyze functions into easy-to-analyze functions with exactly the same optima.

A ***positive monotonic function*** is a function that increases throughout its domain. A positive monotonic function can be either strictly increasing or non-decreasing. Algebraically, a non-decreasing monotonic function has the property that for all x, y such that $x \leq y$ one has $f(x) \leq f(y)$. Replacing these inequalities with strict inequalities yields the definition of a strictly increasing monotonic function.

A ***positive monotonic transformation*** is achieved by plugging the function you want to analyze into any positive monotonic function of your choice.

Negative monotone functions and transformations are defined analogously, and if the context is clear, such functions (transformations) will just be referred to a monotonic functions (transformations).

Key result: Any monotonic transformation of a function has the same optima as the original function!

A characteristic of functions is called **ordinal** if every monotonic transformation of a function with this characteristic still has this characteristic. On the other hand, **cardinal** properties are not preserved by monotonic transformations. Importantly, utility is an ordinal concept:

- For example, let $u(x, y) \in \mathbb{R}^2_+$ be a utility function and let $v(x, y) = u(x, y) + 1$ be another utility function $\Rightarrow$ same set of indifference curves $\Rightarrow$ same preferences.
- The desirable properties of concavity/convexity are not applicable when working with utility functions because they are cardinal.

Examples:

Economic application: The Cobb-Douglas utility function has the form

$$u(x_1, x_2) = x_1^{\alpha} x_2^{1-\alpha}.$$

Check that by taking the natural log of the function (i.e. take $ln\ (u(x_1, x_2))$), the optima of the two functions are the same under the constraint $x_1 + x_2 \leq 100$.

Econometric example: The likelihood function takes the form,

$$\mathcal{L}(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta),$$

and the loglikelihood takes the form

$$\ell(\theta|x) = \log \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} \log f(x_i|\theta).$$

Test yourself again:

$$\max_{x_1, x_2} f(x) = e^{\sqrt{x_1 x_2}} \text{ s.t. } x_1 + 4x_2 = 16.$$

1.8 Quasiconcavity and Quasiconvexity

Quasiconcavity and quasiconvexity are concepts that apply to functions. Here we present graphical and algebraic definitions of quasiconcavity and quasiconvexity.

Quasiconcavity:

First, let f be a function of many variables (the vector x, say) defined on the set S. Then for any $a \in \mathbb{R}$, the set

$$P_a^+ = \{x \in S : f(x) \geq a\}$$

is called the ***upper level set*** of f for a.

The function f of many variables defined on a convex set S is ***quasiconcave*** if every upper level set of f is convex. (That is, $P_a^+ = \{x \in S : f(x) \geq a\}$ is convex for every value of a.)

We also have an equivalent algebraic representation: A function f is ***quasiconcave*** if and only if, for every pair of distinct points u and v in the domain of f, and for $\theta \in (0,1)$,

$$f(v) \geq f(u) \implies f(\theta u + (1-\theta)v) \geq f(u).$$

If the second inequality is strict, then f is ***strictly quasiconcave***.

Quasiconvexity:

Again, let f be a function of many variables (the vector x, say) defined on the set S. Then for any $a \in \mathbb{R}$, the set

$$P_a^- = \{x \in S : f(x) \leq a\}$$

is called the ***lower level set*** of f for a.

The function f of many variables defined on a convex set S is ***quasiconvex*** if every lower level set of f is convex. (That is, $P_a^- = \{x \in S : f(x) \leq a\}$ is convex for every value of a.)

Checking quasiconcavity and quasiconvexity:

To see whether a function is quasiconcave or quasiconvex one can examine the level sets of the function directly. Alternatively if the function is differentiable (twice differentiable in one case), two helpful propositions can determine quasi-concavity and quasiconvexity.[6]. We present them in the supplementary

[6] For more information, see Osborne, chapter 3.4 and the cites therein.

section on optimization, since they do not provide much intuition and require material from later chapters of these notes.

Useful results:

- Every concave (convex) function is quasiconcave (quasiconvex).
- If $f(x)$ is quasiconcave, then $-f(x)$ is quasiconvex.
- Any monotonic transformation of a quasiconcave (quasiconvex) function is also quasiconcave (quasiconvex).[7]

Key result:

Knowing whether a function is *strictly* quasiconcave or *strictly* quasiconvex implies that any unconstrained local optima are also global optima. There is thus no need to check second-order conditions if $f(x)$ is strictly quasiconcave or strictly quasiconvex, for finding the FOC of a strictly quasiconcave (strictly quasiconvex) function finds a global maximum (minimum).

[7] Unlike concavity and convexity, quasi-concavity and quasi-convexity retain their properties of quasiconcavity/quasiconvexity when they are monotonically transformed. This is a useful property for certain objective functions to have. Consider utility functions, which tend to be ordinal concepts, and thus their interpretation should be immune to monotonic transformations. For, with ordinal utility, we only care about the preference ordering of consumption bundles, not the degree to which certain consumption bundles are preferred over one another.

Example: Find the extremum of

$$f(x_1, x_2) = x_1^2 + x_2^2$$

subject to,

$$x_1 + 4x_2 = 2.$$

$$x_1, x_2 \geq 0$$

Draw the objective function:

Now draw the constraint and a few of the objective function's level sets in the $x_1 - x_2$ plane:

Draw the lower and upper level sets for an arbitrary value of $f(x_1, x_2)$. Which set is convex? Is this function quasiconcave or quasiconvex?

What type of extremum are the FOCs sufficient for?

Set up the Lagrangian and solve for the optimum (FOCs only):

Example: Consider the following two Cobb-Douglas functions

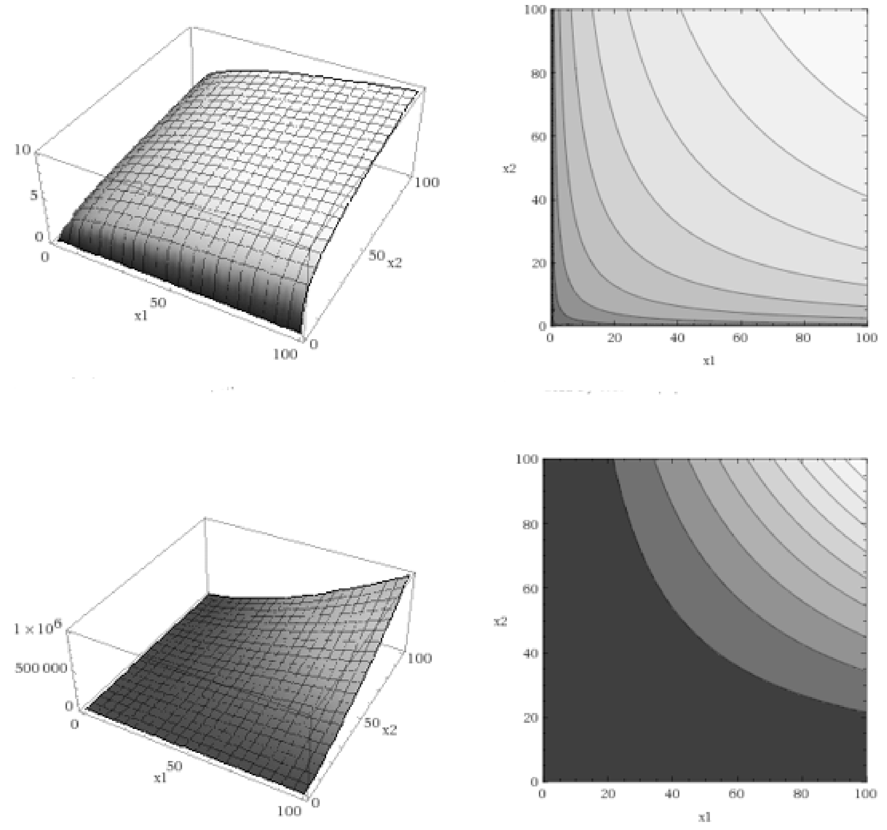


Figure 3: The first function is of the form $f(x) = x_1^{\frac{1}{4}} x_2^{\frac{1}{4}}$; the second function is of the form $f(x) = x_1^{\frac{3}{2}} x_2^{\frac{3}{2}}$.

Note how we can analyze these two different looking Cobb-Douglas functions together as quasiconcave functions.

1.9 Constrained Optimization with Multiple Constraints

This is identical to the case with a single constraint, aside from adding an additional Lagrange multiplier for each constraint. Note though, you can still substitute in constraints where possible.

Consider an objective function $z = f(x, y)$ subject to two constraints, $g(x, y) = c$ and $h(x, y) = d$:

1. Introduce two Lagrange multipliers, λ_1 and λ_2, one for each constraint;
2. Rewrite each constraint with everything on one side of the equation:
$$g(x, y) - c = 0 \text{ and } h(x, y) - d = 0$$
3. Create the Lagrangian function, a modified version of the objective function:
$$\mathcal{L} = f(x, y) - \lambda_1 [g(x, y) - c] - \lambda_2 [h(x, y) - d]$$
4. Solve this unconstrained optimization problem as usual, treating the Lagrange multipliers, λ_1 and λ_2, as additional variables.

5. Check your solution from Step 4 to determine if it's a maximum or minimum. You can use the bordered Hessian approach outlined in Simon and Blume, chapter 19.[8]

Example:

Find the extremum of $z = x^2 + 2xy + yw^2$ subject to

$$2x + y + w^2 = 24$$

and

$$x + w = 8.$$

[8] We will not cover this topic during math camp, but it is important for you to know where to look if you encounter the need to check your SOCs.

2 Comparative Statics (Day 2)

Sources for this section can be found in Chiang and Wainwright, chapters 8.1-8.7, and Simon and Blume, chapters 14.4, 15.1 - 15.4.

2.1 Total Differentiation

We reviewed how to compute how much a function changes when one variable changes – that's a partial derivative. How does one compute how much a function changes when multiple variables change at the same time?

To answer this question, we compute the ***total differential***.

To understand the total differential, consider the function $f : \mathbb{R}^2 \to \mathbb{R}$. Consider now the tangent plane to f at some point in the domain of f, (x^*, y^*), say:

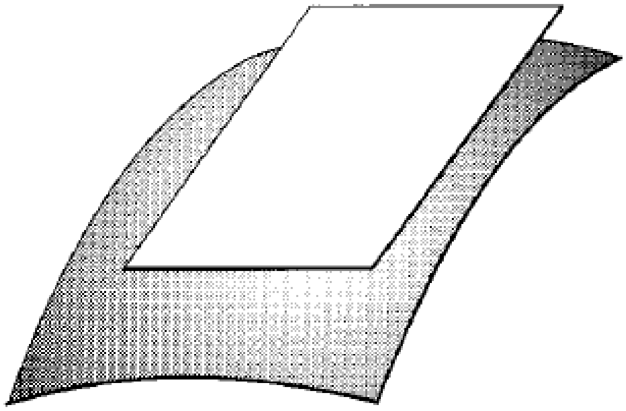


Figure 4: The tangent plane to the function f at (x^*, y^*).

We use the following notation to depict changes on the tangent plane:

$$dx = \Delta x,\, dy = \Delta y, \text{ and } df = \frac{\partial f}{\partial x}(x^*, y^*)dx + \frac{\partial f}{\partial y}(x^*, y^*)dy.$$

These variations on the tangent plane are called ***differentials***. We have that the change on the function f is approximately the change df *on the tangent plane*, given Δx and Δy. The above expression for df in terms of dx and dy is called the ***total differential*** of f at (x^*, y^*).

Examples:

Compute the total differential of the function $f(x,y) = x^2 + xy + y^2$.

Compute the total differential of the savings function $S = S(y,i)$ where y is income, i is the interest rate, and S is savings.

Compute the total differential of the utility function $U = U(x_1, x_2, \ldots, x_n)$.

2.2 Total Derivatives

How does one compute the derivative of a function when the arguments of that function are related?

To answer this question, compute the ***total derivative***.

To understand the total derivative, consider a function $f(x_1(t), \ldots, x_n(t)) : \mathbb{R}^n \to \mathbb{R}$. As the notation suggests, let us assume that the variables $x_1, \ldots, x_n$ themselves are related through t. Thus, to see how the function changes with t, we compute

$$\frac{df}{dt} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{dx_i}{dt}.$$

This is what is known as the ***total derivative*** of f with respect to t. Notice that it is nothing but a glorified application of the chain rule.

Examples:

Compute the total derivative of the function $f(x,y) = x^2 + xy + y^2$ with respect to x.

Compute the total derivative, $\frac{dS}{di}$, of the function $S = S(y,i)$, where $y = f(i)$.

Consider the production function $F = F(K,L,t)$, where L denotes labor, K denotes capital, and t denotes time. Assume that capital and labor both vary with time. Compute the rate of change of output with respect to time.

Given the production function $F(K, L)$, show that, along an isoquant,[9] the marginal rate of technical substitution is

$$MRS_K^L = \frac{dL}{dK} = -\frac{\frac{\partial F}{\partial K}}{\frac{\partial F}{\partial L}}$$

[9] An isoquant is the set of production inputs that produce the same level of output.

2.3 Implicit Functions and their Derivatives

We first introduce some definitions:

- An ***implicit equation*** is of the form $F(y, x_1, \dots, x_n) = 0$. For each set of x, there may exist multiple values for y that solve the equation. e.g. consider the unit circle:
$$x^2 + y^2 - 1 = 0.$$
- An ***explicit function*** is a function of the general form $y = F(x_1, ..., x_n)$ where the endogenous (dependent) variables (e.g., y) are solved in terms of the exogenous (independent) variables (e.g. x). For each set of exogenous variables, there exists a *unique* value for each endogenous variable, e.g.
$$y = +\sqrt{1 - x^2} \qquad\qquad y = -\sqrt{1 - x^2}$$
- If for each exogenous variable $(x_1, ..., x_n)$ an implicit equation determines a corresponding value y, then we say that the implicit equation defines the endogenous variable y as an ***implicit function*** of the exogenous variables, e.g.:
$$y^2 + f(x)^2 - 1 = 0$$
Just because we can write down an explicit equation, say $F(x, y) = c$ does not mean that this equation automatically defines y as a function of x. Consider the implicit equation defining the unit circle, $x^2 + y^2 - 1 = 0$: When $x > 1$, there is no y satisfying this implicit equation. However, typically think of defining implicit functions at a specific solution to an implicit equation (e.g., $x = 0, y = 1$) and then ask questions about the implicit function in the neighborhood of that solution.

More examples:

- $y = 3x^4$ is an explicit function
- $y = f(x)$ is an explicit function
- The implicit equation $4x + 2y - 5 = 0$ defines y as an implicit function of x
- $y^2 - 5xy + 4x^2 = 0$ defines y as multi-valued implicit function of x

- $y^5 - 5xy + 4x^2 = 0$ is an implicit equation—which cannot be solved into an explicit function—that defines y as a function of x *for certain values of* x (e.g., when $x = 0, y = 0$)
- $F(y, x_1, \ldots, x_n) = 0$ is an implicit equation that may or may not implicitly define an implicit equation

Question: Is $F(y, x) = x^2 + y^2 - 1 = 0$ an implicit function?

Implicit function theorem:

Given an *implicit equation* $F(y, x_1, \ldots, x_n) = 0$, if:

1. F has continuous partial derivatives, and
2. at a point $(y^0, x_1^0, \ldots, x_n^0)$ satisfying the equation $F(y, x_1, \ldots, x_n) = 0$,

$$\frac{\partial F}{\partial y}(y^0, x_1^0, \ldots, x_n^0) \neq 0,$$

then there exists an n-dimensional neighborhood of $(x_1^0, \ldots, x_n^0)$ in which y is an implicitly defined function of the variables $(x_1, \ldots, x_n)$.

Note that while this is a sufficient condition, it is not necessary; consider the implicit equation $y^3 - \alpha = 0$.

Now, test whether the conditions of the Implicit Function Theorem hold for $F(y, x) = x^2 + y^2 - 9 = 0$. When does the implicit equation (implicitly) define y as a function of x?

Implicit Function Rule:

Now we know when an implicit equation defines a one-to-one relationship between its independent and dependent arguments. With such a one-to-one mapping, it now makes sense to consider how the function's values (equivalently, its output y) changes when its arguments change.

The ***implicit function rule*** states:

Given an *implicit equation* $F(y, x_1, \ldots, x_n) = 0$, if an implicit function exists, the the partial derivatives are given by:

$$\frac{\partial y}{\partial x_i} = -\frac{\frac{\partial F}{\partial x_i}}{\frac{\partial F}{\partial y}}$$

Example:

Compute the partial derivative $\frac{\partial y}{\partial x}$ of the implicit equation $F(y, x) = x^2 + y^2 - 9 = 0$ and evaluate it at $(0, 3)$.

2.4 Comparative Statics

As is hinted by the name, ***comparative statics*** typically concern how an optimum or an equilibrium condition changes when an underlying, exogenous parameter changes.

To see how equilibrium conditions or optimal choices change when such underlying, exogenous parameters change, we can employ the tools of today's subsection. We'll consider two examples to illustrate the concept of comparative statics.

Example: Tax incidence in a perfectly competitive market.

Market equilibrium is given by:

$$\begin{aligned} Q^d &= D(p+t) \\ Q^s &= S(p) \\ Q &= Q^d = Q^s \end{aligned}$$

where Q^d is quantity, Q^s is quantity supplied, p is price, and t is a tax levied on consumers. Determine how a change in the tax, t, affects the equilibrium price p. Assume consumers respond identically to changes in prices and taxes. Try rewriting your solution in terms of elasticities of demand and supply. If you use any theorems, note what assumptions you have made.

Example: The optimal forest rotation problem.

Imagine we have a forest whose value at time T is given by the function $V(T)$, where $V''(T) < 0$. The value function $V(T)$ is maximized when $V'(T) = 0$, but if you can cut down and sell the trees, put your money in the bank, and earn interest rate δ, then you wouldn't necessarily want to let the trees grow until they reach their maximum value. Assume the value function is globally nonnegative. Your goal is to maximize the net present value of the forest:

$$\max_T f(T) = e^{-\delta T} V(T)$$

1. Derive an expression for the optimal harvest time T^*, taking δ as fixed. Note any assumptions you make.

2. Calculate $\frac{dT^*}{d\delta}$, which shows how the optimal harvest time changes if the interest rate changes.

3 Linear Algebra (Day 2)

Useful sources for this section include Chiang and Wainwright, chapters 4-5, and Simon and Blume, chapters 8-11.

Helpful additional resources include The Matrix Cookbook and MIT OpenCourseWare (great video lectures).

3.1 What is a Matrix?

A matrix is a rectangular array of numbers, variables, or functions. Matrices are used to efficiently store and manipulate information. Examples:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{33} \\ a_{41} & a_{43} \end{bmatrix} \qquad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

Elements: Individual terms in a matrix (e.g., a_{11}).

Equality: Two matrices are equal if and only if they have the same dimension and identical elements: $A = B \iff a_{ij} = b_{ij}, \forall i, j$.

Dimensions of a matrix: (number of rows) x (number of columns)

What are the dimensions of A?

What are the dimensions of B?

In what follows, the dimensions of a matrix are denoted by $n \times m$, where n is the number of rows and m the number of columns.

3.2 Important Classes of Matrices

Scalar: a matrix with one row and one column

$$\begin{bmatrix} 7 \end{bmatrix}$$

Vector: a matrix with either one row or one column. Here is a column and a row vector. When left unspecified, a vector typically refers to a column vector.

$$A = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}$$

$$B = \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix}$$

Square matrix: a matrix with the same number of rows as columns ($N = M$)

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

Symmetric matrix: a matrix that equals its own transpose

$$\begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix}$$

Identity matrix: a square diagonal matrix with all diagonal elements equal to one

$$I_1 = \begin{bmatrix} 1 \end{bmatrix} \qquad I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Null matrix: a matrix of zeros; often denoted $\mathbf{0}_{(n \times m)}$

$$\mathbf{0}_{(2\times3)} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Idempotent matrix: a matrix that, when multiplied by itself, equals itself

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

Positive-definite matrix: an $n \times n$ matrix is **positive definite** if:

$$v'Av > 0, \quad \forall v \in \mathbb{R}^n, \quad v \neq \mathbf{0}_{n\times1}$$

Similarly, an $n \times n$ matrix is **positive semidefinite** if:

$$v'Av \geq 0, \quad \forall v \in \mathbb{R}^n$$

Flip inequalities for **negative definite** and **negative semidefinite**.

Applications:

- Statistics: All variance-covariance matrices are positive semidefinite.
- Optimization: The definiteness of a symmetric matrix is important for multivariate optimization. These properties are roughly analogous to concavity and convexity; the Second Order Conditions of an optimization problem can be checked by determining whether the matrix of second derivatives (Hessian) is positive or negative definite.[10]

[10] See Simon and Blume Chapters 16-18 for more information.

3.3 Basic Matrix Operations

Transposition: a matrix with its rows and columns switched. The transpose of a matrix A is denoted by A' or A^T.

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{bmatrix}$$

$$A' = \begin{bmatrix} a_{11} & a_{21} & \ldots & a_{n1} \\ a_{12} & a_{22} & \ldots & a_{n2} \\ \ldots & \ldots & \ldots & \ldots \\ a_{1m} & a_{n2} & \ldots & a_{nm} \end{bmatrix}$$

Addition and Subtraction: matrices must be *conformable for addition*—they must have equivalent dimensions. If the matrices are conformable, simply add or subtract the corresponding elements of the matrices.

$$A + B = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \ldots & a_{1m} + b_{1m} \\ a_{21} + b_{21} & a_{22} + b_{22} & \ldots & a_{2m} + b_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} + b_{n1} & a_{n2} + b_{n2} & \ldots & a_{nm} + b_{nm} \end{bmatrix}$$

Scalar multiplication: multiply each of the elements of the matrix by the scalar.

For $c \in \mathbb{R}$, we have,

$$cA = \begin{bmatrix} ca_{11} & ca_{12} & \dots & ca_{1m} \\ ca_{21} & ca_{22} & \dots & ca_{2m} \\ \dots & \dots & \dots & \dots \\ ca_{n1} & ca_{n2} & \dots & ca_{nm} \end{bmatrix}$$

Dot (scalar) product of two vectors: to multiply a $1 \times n$ row vector by a $n \times 1$ column vector, multiply the corresponding elements, and add up the products. Note the length of the two vectors must be the same for the dot product to be defined.

Let $v_{n\times 1}$ and $u_{n\times 1}$, then

$$v \cdot u = v'u = \sum_{i=1}^{n} v_i u_i$$

The dot product is a special case of the inner product, specific to Euclidean space.

Matrix Multiplication

Step 1: Are the matrices *conformable for multiplication*? The number of columns of the first matrix must equal the number of rows of the second matrix.
Exception: a scalar (1×1 matrix) and matrix of any size can be multiplied together.

Step 2: Determine the dimensions of the product. If A is $n \times m$ and B is $m \times q$, then $C = AB$ is $n \times q$.

Step 3: Multiply! The element c_{ij} of the product matrix C is equal to the dot product of row i of matrix A and column j of matrix B.

That is, AB = $[c_{ij}]$, where $c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{im}b_{mj}$, for every ith row and every jth column.

Practice Problems

Example: Addition:

$$\begin{bmatrix} 1 & 3 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 7 & 5 \\ 2 & 1 \end{bmatrix} =$$

Example: Scalar multiplication:

$$4\begin{bmatrix} 1 & 3 & 5 \\ -1 & -8 & 10 \\ -7 & -5 & 13 \end{bmatrix} =$$

Example: Matrix multiplication:

$$\begin{bmatrix} 2 & 0 & -1 & 1 \\ 1 & 2 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 5 & -7 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \\ 2 & 0 & 0 \end{bmatrix} =$$

3.4 Rules of Matrix Operations

Commutative Law of Addition

$$A + B = B + A$$

(No) Commutative Law of Multiplication

$$AB \neq BA$$

Associative Law of Addition

$$(A + B) + C = A + (B + C)$$

Associative Law of Multiplication

$$(AB)C = A(BC)$$

Additive Inverse For any matrix A, define $-A = [-a_{ij}]$. Then

$$A + (-A) = \mathbf{0}$$

Distributive Law

$$A(B + C) = AB + AC$$

No "Zero Product Rule"

$$AB = \mathbf{0} \not\Rightarrow A = \mathbf{0} \text{ or } B = \mathbf{0}$$

Properties of the Transpose

$$\begin{aligned} (A^T)^T &= A \\ (A + B)^T &= A^T + B^T \\ (ABC)^T &= C^T B^T A^T \end{aligned}$$

Properties of the Identity Matrix . For the $n \times n$ identity matrix I and any $n \times n$ matrix A,

$$\begin{aligned} AI = IA &= A \\ I(I) &= I \end{aligned}$$

Properties of the Null Matrix. For any matrix A which is $m \times n$,

$$\begin{aligned} A + \mathbf{0}_{(m \times n)} &= A \\ A\mathbf{0}_{(n \times p)} &= \mathbf{0}_{(m \times p)} \end{aligned}$$

3.5 Basic Properties of Matrices

3.5.1 Matrix Rank

Row (column) rank: the number of linearly independent row (column) vectors, where a row (column) vector is linearly independent if it cannot be computed as a linear function of the other row (column) vectors.

Rank: the number of linearly independent rows or columns, or the row rank or the column rank (the two are always equal). The rank of a matrix is denoted rank(A).

Determining the rank of a matrix:

For small matrices, it is often possible to determine the rank of a matrix by inspection, comparing rows or columns to test for linear dependence. For instance, find

$$\text{rank}\begin{bmatrix} 2 & -4 \\ -1 & 2 \end{bmatrix} =$$

Matrix rank can be computed using **Gauss-Jordan elimination** – elementary row operations: addition, subtraction, and scalar multiplication – to reduce a matrix to **row echelon form**, in which each row has more leading zeros than the row preceding it. This can be used to determine the matrix rank, which is also defined as the number of nonzero rows of a matrix in row echelon form.

Gauss-Jordan elimination operations:

- Swapping two rows
- Multiplying a row by a nonzero number
- Adding a multiple of one row to another

Example: Use Gauss-Jordan elimination to verify your solution by inspection for

$$\text{rank}\begin{bmatrix} 2 & -4 \\ -1 & 2 \end{bmatrix} =$$

Properties of the matrix rank For an $M \times N$ matrix A,

- $\text{rank}(A) \leq \min\{M, N\}$. If $\text{rank}(A) = \min\{M, N\}$, matrix A is of **full rank**.
- If $M = N$, A is **invertible** if and only if A has full rank.
- If the product AB is well–defined, $\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$.
- Subadditivity: If the sum $A + B$ is well–defined, $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$.

Examples: Find the maximum possible rank of each matrix based solely on dimensions.

$$\text{rank}\begin{bmatrix} 2 & -4 & 3 \\ -1 & 2 & 0 \end{bmatrix} \leq$$

$$\text{rank}\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix} \leq$$

$$\text{rank}\left(\begin{bmatrix} a & b \\ d & e \end{bmatrix}\begin{bmatrix} 2 & -4 & 3 \\ -1 & 2 & 0 \end{bmatrix}\right) \leq$$

Examples: Use Gauss-Jordan elimination to compute

$$\text{rank}\begin{bmatrix} 1 & 6 & -7 & 3 \\ 1 & 9 & -6 & 4 \\ 1 & 3 & -8 & 4 \end{bmatrix} =$$

Does this matrix have full rank?

Is this matrix invertible?

3.5.2 Matrix Determinant

Like rank, the determinant is a real-valued function of matrices. The determinant of the $n \times n$ matrix A is denoted $\det(A)$. The key takeaway is the determinant is used to test whether a matrix is of full rank (i.e., $\det(A) \neq 0 \Leftrightarrow \text{rank}(A) = N$), which also serves as a test of invertibility. In practice, you can use a computer to calculate the determinant of large matrices. Here are three handy rules you can use for some matrices:

Calculating the determinant of a 2 x 2 matrix

$$\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} =$$

Calculating the determinant of a 3 x 3 matrix

$$\det\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = a_{11}\det\begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} - a_{12}\det\begin{pmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{pmatrix} + a_{13}\det\begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}$$

Simplify the determinant using the formula for the determinant of a 2×2 matrix:

Calculating the determinant of a triangular matrix:

The determinant of a lower-triangular, upper-triangular, or diagonal matrix is simply the product of its diagonal entries: $\det(A) = \prod_{i=1}^{N} a_{ii}$. Verify this using the formula for a 2×2 matrix:

$$\det\begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix} = \qquad \det\begin{pmatrix} a & b \\ 0 & d \end{pmatrix} = \qquad \det\begin{pmatrix} a & 0 \\ c & d \end{pmatrix} =$$

Properties of the determinant Given an $N \times N$ matrix A, matrix determinants satisfy the following properties:

- $\det(A) \neq 0 \Leftrightarrow \text{rank}(A) = N$.
- If $\det(A) \neq 0$, then $\det(A^{-1}) = \dfrac{1}{\det(A)}$.
- For any scalar $\alpha \neq 0$, $\det(\alpha A) = \alpha^N \det(A)$.
- If A is diagonal or triangular, $\det(A) = \prod_{n=1}^{N} a_{nn}$.
- $\det(A^T) = \det(A)$.

3.5.3 Matrix Trace

The trace is another real–valued function that can be defined only on square matrices, and corresponds to the sum of the terms along the main diagonal. Given an $N \times N$ matrix A, the trace of A is given by

$$\text{tr}(A) = \sum_{n=1}^{N} a_{nn}.$$

Given an $N \times N$ matrix A, matrix trace satisfy the following properties:

- Linearity: $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$ and $\text{tr}(cA) = c\text{tr}(A)$.
- Invariance under Transposition: $\text{tr}(A) = \text{tr}(A^T)$.

Examples: Compute the trace of the following matrices:

$$\text{tr}\left(\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}\right) =$$

$$\text{tr}\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} e & f \\ g & h \end{bmatrix}\right) =$$

$$\text{tr}\left(3\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}\right) =$$

3.6 Matrix Inversion

Inverse of a matrix: Loosely speaking, the inverse of a matrix is a generalization of the scalar inverse. Formally, the inverse of a matrix A is a matrix A^{-1} such that $AA^{-1} = A^{-1}A = I$. A few notes:

- Only square matrices have inverses.
- A matrix possessing an inverse is also referred to as nonsingular, nondegenerate, or invertible.
- If an inverse exists, then it is unique.

The following definitions are equivalent to an $N \times N$ matrix A being invertible:

- $\text{rank}(A) = N$.
- The rows (or columns) of A are linearly independent.
- $\det(A) \neq 0$.
- A^T is invertible.

Common use: Solve systems of linear equations.

Properties of Inverses:

$$\begin{aligned}(A^{-1})^{-1} &= A \\ (A^T)^{-1} &= (A^{-1})^T \\ (AB)^{-1} &= B^{-1}A^{-1} \\ (kA)^{-1} &= k^{-1}A^{-1} \text{ for nonzero scalar } k\end{aligned}$$

Handy fact: The inverse of a diagonal matrix is a matrix of the (scalar) inverses of the diagonal elements.

Matrix Inversion: Summary of Three Procedures:

Formula for 2 x 2 Matrix

$$A^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Example: Compute

$$\begin{bmatrix} 1 & 3 \\ 1 & 4 \end{bmatrix}^{-1} =$$

Gauss-Jordan Elimination

1. Does the inverse exist? Check by verifying the matrix is full rank (e.g., is the determinant nonzero?).
2. Write the matrix in augmented form with the identity matrix.

   $$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \Rightarrow \quad \left[\begin{array}{cc|cc} a & b & 1 & 0 \\ c & d & 0 & 1 \end{array}\right]$$

3. Perform row operations to reduce the original matrix to the identity matrix.

   This works for any square matrix, regardless of size.

Example: Compute

$$\begin{bmatrix} 1 & 3 \\ 1 & 4 \end{bmatrix}^{-1} =$$

Matrix Inversion using Minors, Cofactors, and Adjoint

1. Does the inverse exist? Check by calculating the determinant.
2. Create the matrix of minors: for each element of the matrix, ignore values in the current row and column, and calculate the determinant of the remaining values.
3. Create the cofactor matrix: change the sign of every other cell in the matrix of minors according to the rule:

$$\sigma(i+j) = \begin{cases} 1, & \text{if } i+j \text{ is even} \\ -1, & \text{if } i+j \text{ is odd,} \end{cases}$$

4. Create the adjoint matrix, which is just the transpose of the cofactor matrix.
5. Divide the adjoint matrix by the determinant of the original matrix to find the inverse matrix.

Example: Use this method to verify

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Practice Problems:

Example: Multiply

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \text{and} \qquad A^{-1} = \begin{bmatrix} \frac{d}{ad-bc} & \frac{-b}{ad-bc} \\ \frac{-c}{ad-bc} & \frac{a}{ad-bc} \end{bmatrix}$$

to verify the property $AA^{-1} = I$.

Example: Determine whether the following matrix is invertible by computing its determinant. If it is, use the formula for a 2×2 matrix to compute:

$$\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}^{-1} =$$

Example: Determine whether the following matrix is invertible by computing its determinant. If it is, use Gauss-Jordan elimination to compute:

$$\begin{bmatrix} 2 & 1 \\ -4 & -2 \end{bmatrix}^{-1} =$$

Example: Determine whether the following matrix is invertible by computing its determinant. If it is, use Gauss-Jordan elimination to compute:

$$\begin{bmatrix} 4 & 5 \\ 2 & 4 \end{bmatrix}^{-1} =$$

3.7 Solving Systems of Linear Equations

Suppose we've been given a system of linear equations:

$$\begin{aligned} 4x + 3y &= 28 \\ 2x + 5y &= 42 \end{aligned}$$

Solutions methods:

1. Regular old algebra.
2. Linear algebra:
   (a) Rewrite the system in matrix form: $Ax = b$.
   (b) Invert the matrix A. *Note: If the matrix is invertible, the system of equations has a unique solution.*
   (c) Use matrix multiplication to solve for x using the formula $x = A^{-1}b$.[11]

[11] To see this, pre-multiply the original equation by A^{-1}: $A^{-1}Ax = A^{-1}b \Rightarrow Ix = A^{-1}b \Rightarrow x = A^{-1}b$.

Practice Problems:

Example: Invert the coefficient matrix (A) to solve the following system of equations:

$$\begin{aligned} 2x_1 + x_2 &= 5 \\ x_1 + x_2 &= 3. \end{aligned}$$

3.8 Matrix Calculus

Below are some formulas that come in handy in econometrics and micro theory:

- The gradient of a scalar function $f : \mathbb{R}^N \to \mathbb{R}$ is given by

$$\nabla f \equiv \frac{\partial f}{\partial x} = \left[\begin{array}{ccc} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_N} \end{array} \right].$$

  Note: This is a special case of the Jacobian, which generalizes the gradient of scalar functions to vector-valued functions.

- The tangent vector to a curve $x : \mathbb{R} \to \mathbb{R}^N$ is given by

$$\frac{\partial x(t)}{\partial t} = \left[\begin{array}{c} \frac{\partial x_1(t)}{\partial t} \\ \vdots \\ \frac{\partial x_N(t)}{\partial t} \end{array} \right].$$

- Given the linear function $z = c^T x = x^T c$,

$$\frac{\partial z}{\partial x} = c$$

- Given a matrix A and the linear function $y = Ax$,

$$\begin{aligned} \frac{\partial y}{\partial x} &= A^T \\ \frac{\partial (x^T A x)}{\partial x} &= (A + A^T)x \\ \frac{\partial (x^T A x)}{\partial A} &= xx^T \end{aligned}$$

- Given $x, v \in \mathbb{R}^k$,

$$\frac{\partial x^T v}{\partial v} = \frac{\partial v^T x}{\partial v} = x$$

3.9 Common Matrices in Economics and Econometrics

Budget constraint: $p \cdot x = w$ or $p'x = w$

Quadratic matrix: $X'X$—analogous to squaring a scalar; produces a symmetric matrix.

Jacobian: the matrix of all first-order partial derivatives.

Hessian: the matrix of all second-order partial derivatives.

Note: this is a symmetric matrix (recall the order of partial differentiation does not matter).

Sample mean: If X is a $n \times 1$ random matrix, the sample mean equals $\frac{1}{n}X'\mathbf{1}_n$.

Sample variance-covariance matrix: If X is a $n \times m$ random matrix, the sample variance-covariance matrix is given by:

$$S = \frac{1}{n-1}X'\left(I_n - \frac{1}{n}\mathbf{1}_{n\times n}\right)X$$

Practice Problems:

Example: What are the Jacobian and Hessian for the function $f(x, y) = xy$?

Example: What are the Jacobian and Hessian for the function $f(x, y) = 4x^2y - 3xy^3 + 6x$?

3.10 Application: The Least-Squares Estimator

Goal of regression: To find the relationships between an outcome variable and one or more explanatory variables.

Example: What are the determinants of income?

Assumption: Income is a linear function of age, race, gender, education, and some other unobservable factors:

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 Age_1 + \beta_2 Race_1 + \beta_3 Gender_1 + \beta_4 Education_1 + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 Age_2 + \beta_2 Race_2 + \beta_3 Gender_2 + \beta_4 Education_2 + \varepsilon_2 \\
y_3 &= \beta_0 + \beta_1 Age_3 + \beta_2 Race_3 + \beta_3 Gender_3 + \beta_4 Education_3 + \varepsilon_3 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 Age_n + \beta_2 Race_n + \beta_3 Gender_n + \beta_4 Education_n + \varepsilon_n
\end{aligned}$$

Put in matrix form: Construct matrices Y, X, β, and ε:

Question: How do we estimate β, the vector of coefficients?

Ordinary Least Squares (OLS): minimize the sum of squared residuals.

1. For any β we select (call it b), we can construct a vector of residuals, e.

2. Choose the vector b that minimizes the sum of squared residuals, $e'e$.

4 Probability and Statistics (Day 3)

Useful sources for this section include Simon and Blume, Appendix 5, and Wooldridge (2000), appendices B and C.

4.1 Random Variables and Distributions

4.1.1 Random Variables

A random variable is a function that maps outcomes from a sample space to real numbers.[12] Intuitively, random variables assign a number to each possible outcome of an uncertain event. Random variables:

- are usually denoted by an upper-case letter, and
- can be discrete (e.g., heads or tails) or continuous (e.g., time to event).

Random variables are characterized by a **probability density function (pdf)**, $f_X(x)$, and a **cumulative distribution (density) function (cdf)**, $F_X(x)$:

$$\begin{aligned} f_X(x) &= \Pr(X = x) \\ F_X(x) &= \Pr(X \leq x) \end{aligned}$$

For discrete random variables, the pdf is also referred to as a **probability mass function**.

The **cdf** F_X has the following properties:

1. for $x_1 \leq x_2$,
$$F_X(x_2) - F_X(x_1) = \Pr(x_1 < X \leq x_2)$$
2. $\lim_{x \to -\infty} F_X(x) = 0, \lim_{x \to \infty} F_X(x) = 1$
3. $F_X(x)$ is non-decreasing
4. $F_X(x)$ is right continuous: $\lim_{x \to x_0^+} F_X(x) = F_X(x_0)$

If F_x is constant except at a countable number of points (i.e., F_x is a step function), then we say that X is a **discrete random variable**. Note the following about discrete random variables:

- The size of the jump in the cdf at x_i is the probability that X takes on the value x_i:
$$p_i = F_X(x_i) - \lim_{x \to x_i^-} F_X(x) = \Pr(X = x_i)$$

[12] The sample space is the set of all possible outcomes of a random variable.

- The cdf or probability mass function (pmf) of X is defined as:

$$f_X(x) = \begin{cases} p_i & \text{if } x = x_i, i = 1, 2, ... \\ 0 & \text{otherwise} \end{cases}$$

- We can write

$$\Pr(x_1 < X \leq x_2) = \sum_{x_1 < x \leq x_2} f_X(x)$$

If F_x can be written as

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

where $f_X(x)$ satisfies

$$f_X(x) \geq 0, \forall x \in \mathbb{R}$$
$$\int_{-\infty}^{\infty} f_X(t)dt = 1$$

then we say that X is a **continuous random variable**. Continuous RVs have a nice relationship between the cdf and pdf; by the Fundamental Theorem of Calculus, where f_X is continuous:

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Some (potentially) helpful notation: We call

$$S_X = \{x : f_X(x) > 0\}$$

the **support** of X. Note that for $x_2 \geq x_1$,

$$\begin{aligned} \Pr(x_1 < X \leq x_2) &= F_X(x_2) - F_X(x_1) \\ &= \int_{x_1}^{x_2} f_X(t)dt \end{aligned}$$

Interestingly, for a continuous RV X,

$$\Pr(X = x) = 0$$

due to the infinitely large support of X. As a result, do not interpret the pdf of a continuous RV as expressing a probability ($f_X(x) \neq \Pr(X = x)$); the proper interpretation is that $f_X(x)$ expresses the probability that X falls within some small interval $(x, x + \Delta x)$.

4.1.2 Joint Distributions and Independence

Let X, Y be two scalar random variables; the **joint cumulative distribution function** of X, Y is

$$F_{X,Y}(x, y) = \Pr(X \leq x, Y \leq y)$$

For discrete random variables X, Y, the joint cdf is

$$F_{X,Y}(x, y) = \sum_{u \leq x} \sum_{v \leq y} f_{X,Y}(u, v)$$

where $f_{X,Y} = \Pr(X = x, Y = y)$ is the joint pmf of X, Y.

For continuous random variables X, Y, the joint pdf is

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v) dv du$$

where again $f_{X,Y}(x, y)$ is the joint pdf of X, Y.

Suppose X and Y are discrete random variables. Then X and Y are said to be independent if and only if

$$\Pr(X = x, Y = y) = \Pr(X = x) \Pr(Y = y)$$

Suppose X and Y are random variables. Then X and Y are said to be independent if and only if

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

Suppose X and Y are random variables. We say that X and Y are **independent and identically distributed (i.i.d. or IID)** if X and Y are independent

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

and have the same distribution

$$f_X(y) = f_Y(y) \text{ and } F_X(x) = F_Y(y)$$

. This concept, *i.i.d.*, will come up frequently when doing asymptotic statistics in econometrics.

4.1.3 Marginal Distributions

Marginal, or individual, distributions can be derived from the joint distribution for two variables. From the joint cdf of (X, Y), we can recover the **marginal cdf** of X

$$\begin{aligned} F_X(x) &= \Pr(X \leq x) \\ &= \Pr(X \leq x, Y \leq \infty) \\ &= \lim_{y \to \infty} F_{X,Y}(x,y) \end{aligned}$$

We can also recover the **marginal pdfs** from the joint pdf using

$$f_X(x) = \sum_y f_{X,Y}(x,y) \text{ if discrete}$$
$$f_X(x) = \int_{S_y} f_{X,Y}(x,y)dy \text{ if continuous}$$

4.1.4 Conditional Distributions

Consider the discrete random variables (X, Y) and let x be such that $f_X(x) > 0$. The **conditional pmf** of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

which satisfies the properties necessary for a well-defined pmf for a discrete random variable:

$$f_{Y|X}(y|x) \geq 0 \qquad\qquad \sum_y f_{Y|X}(y|x) = 1$$

The **conditional cdf** of Y given $X = x$ is then

$$F_{Y|X}(y|x) = \Pr(Y \leq y|X = x) = \sum_{v \leq y} f_{Y|X}(v|x)$$

Consider the analogous case of continuous random variables (X, Y). For any x such that $f_X(x) > 0$, the **conditional pdf** of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

which is well-defined provided that $f_X(x) > 0$. The **conditional cdf** is

$$F_{Y|X}(y|x) = \int_{-\infty}^{y} f_{Y|X}(v|x)dv$$

4.1.5 Transformations of Random Variables

Let X be a random variable with cdf F_X and define the random variable $Y = h(X)$, where h is a one-to-one function whose inverse h^{-1} exists.

If X is discrete and takes on value $x_1, ..., x_n$ then Y is also discrete and takes on the values $y_i = h(x_i)$ for $i = 1, ..., n$. The pmf of Y is given by

$$\begin{aligned} \Pr(Y = y_i) = \Pr(X = h^{-1}(x_i)) \\ f_Y(y_i) = f_X(h^{-1}(y_i)) \end{aligned}$$

If X is continuous and h is increasing, we have that

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) \\ &= \Pr(X \leq h^{-1}(y)) = F_X(h^{-1}(y)) \end{aligned}$$

It follows directly that

$$f_Y(y) = \frac{dF_Y(y)}{dy} = f_X(h^{-1}(y))\frac{dh^{-1}(y)}{dy}$$

When h is decreasing, it follows analogously that

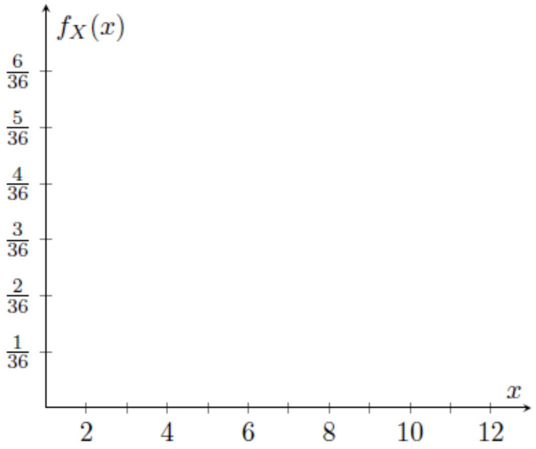
$$f_Y(y) = \frac{dF_Y(y)}{dy} = -f_X(h^{-1}(y))\frac{dh^{-1}(y)}{dy}$$

Problem 1. Suppose you are a basketball player shooting two free throws. You are an 80% free throw shooter. Assuming independence of your free throw attempts, what is the probability that you make both free throws?

You are still a basketball player shooting two free throws. The probability that you make the first free throw is 80%. If you make the first free throw, the probability that you make the second is 85%. If you do not make the first free throw, the probability that you make the second is 70%. What is the probability that you make both free throws?

Problem 2. You throw two fair dice and construct a random variable, X, equal to the sum of the numbers on the two faces. Fill in the table with the values of the pmf of X, $f_X(x)$, considering all possible outcomes of the random variable. Draw a graph of the $f_X(x)$.

x	2	3	4	5	6	7	8	9	10	11	12
$f_X(x)$											



Suppose you throw one die and then the other, and denote the scores by the random variables X and Y. What is the conditional probability that $Y = 4$ given $X = 3$, or $\Pr(Y = 4|X = 3)$? Use the definition of conditional probability.

What about $\Pr(Y = 5|X = 3)$? What is the pmf of the conditional distribution $f_{Y|X}(y|3)$?

Does the conditional distribution depend on the first roll? Why or why not?

Problem 3. Find the marginal probability density functions, $f_X(x)$ and $f_Y(y)$, of the bivariate distribution characterized by the pdf

$$f_{X,Y}(x,y) = (x+y)1_{0\leq x\leq 1}1_{0\leq y\leq 1},$$

where $1_{0\leq x\leq 1}$ represents the indicator function, which takes on the value one for $0 \leq x \leq 1$ and zero otherwise. Are X and Y independent random variables?

4.2 Characteristics of Probability Distributions

4.2.1 Expected Value

The **expected value** can be thought of as a measure of central tendency. In the scalar case, it is a weighted average of all possible realizations of a random variable X (weighted by the probability of each realization). If X is a discrete random variable with J possible realizations, the expected value of X is defined as:

$$\mathbb{E}[X] = \sum_{j=1}^{J} x_j \Pr(X = x_j)$$

If X is a continuous random variable, the expected value of X is defined as:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

Furthermore, the expected value of a function $g(X)$, for discrete X, is given by:

$$\mathbb{E}[g(X)] = \sum_{j=1}^{J} g(x_j) \Pr(X = x_j)$$

Similarly, the expected value of a function $g(X)$, for continuous X, is given by:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Properties of Expected Values

1. For any constant c, $\mathbb{E}[c] = c$.
2. For any constants a and b, $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$.
3. For any constants $\{a_1, a_2, \cdots, a_n\}$ and random variables $\{X_1, X_2, \cdots, X_n\}$,

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i].$$

4. When $a_i = 1 \; \forall i$, Property 3 reduces to

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

Properties 2, 3, and 4 are examples of a general property of the expectation operator known as **linearity of expectation.**

We can use the expectation operator to express probabilities. An **indicator function**, $\mathbb{1}(A)$ is a function that is equal to one if condition A is true and zero otherwise. Consider the case wherre X is a random variable; then

$$\mathbb{1}(X \leq x) = \begin{cases} 1 & \text{if } X \leq x \\ 0 & \text{otherwise} \end{cases}$$

Note that for the continuous case, the following holds:

$$\begin{aligned} \mathbb{E}[\mathbb{1}(X \leq x)] &= \int_{-\infty}^{\infty} \mathbb{1}(X \leq x) f_X(x) dx \\ &= \int_{-\infty}^{x} f_X(x) dx \\ &= F_X(x) = \Pr(X \leq x) \end{aligned}$$

This useful result comes up often in (applied) econometrics!

Here are a few more useful results on the expectations operator:

- Suppose X, Y are random variables with joint density $f_{X,Y}(x,y)$; if $g(x,y) : \mathbb{R}^2 \to \mathbb{R}$, then we have that

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) dy dx$$

- By linearity of expectation, $\forall a, b \in \mathbb{R}$

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

- If X, Y are independent, then for any $h_1(\cdot)$ and $h_2(\cdot)$

$$\mathbb{E}[h_1(X)h_2(Y)] = \mathbb{E}[h_1(X)]\mathbb{E}[h_2(Y)]$$

Problem 4. Calculate the expected value of a roll of one die. If you win \$2 times the face value of your toss, what is your expected winning? *Hint: You just learned a property that can be used to simplify the second part of this question.*

Problem 5. $f_X(x) = 2x$ for $0 \leq x \leq 1$ and $f_X(x) = 0$ otherwise. Calculate $\mathbb{E}[X]$.

Problem 6. Upon graduation, you expect to earn \$50,000 with probability 0.2, \$60,000 with probability 0.5, and \$80,000 with probability 0.3. Your utility function for money is $U = \ln(w)$, where $w =$ your wealth (earnings). Calculate your expected earnings and your expected utility upon graduation.

4.2.2 Conditional Expectations

Suppose Y is a discrete random variable, and X is any random variable. Then the **conditional expectation** of Y given that X equals some value x (i.e., $X = x$) is:

$$\mathbb{E}[Y|X=x] = \sum_{j=1}^{n} y_j \Pr(Y = y_j|X = x)$$

If Y is a continuous random variable, the conditional expectation of Y given $X = x$ is:

$$\mathbb{E}[Y|X=x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy$$

Properties of Conditional Expectations

1. For any function $c(X)$, $\mathbb{E}[c(X)|X = x] = c(x)$.
2. For functions $a(X)$ and $b(X)$, $\mathbb{E}[a(X)Y + b(X)|X] = a(X)\mathbb{E}[Y|X] + b(X)$.
3. If X and Y are independent, then $\mathbb{E}[Y|X] = \mathbb{E}[Y]$.
4. The Law of Iterated Expectations: $\mathbb{E}\left[\mathbb{E}[Y|X]\right] = \mathbb{E}[Y]$.

Exercise: Prove the law of iterated expectations for the continuous case. *Hint: Start with the definition of an expectation and integrate.*

4.2.3 Variance and Standard Deviation

Consider a random variable X. The k**-th moment** of X is defined as $\mathbb{E}[X^k]$.

- The first moment of X is its mean: $\mathbb{E}[X]$
- The k**-th centered moment** of X is $\mathbb{E}[(X - \mathbb{E}[X])^k]$

The second centered moment of X is referred to by a special name: the variance. In contrast to the expectation, which is a measure of central tendency, variance and standard deviation are measures of variability or spread. The **variance** of the random variable X is:

$$\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

Variance is also denoted by σ_X^2 or σ^2. It is frequently useful to express the variance as:

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Exercise: Derive this result.

The **standard deviation** of a random variable X is the (positive) square root of its variance:

$$\text{sd}(X) = \sigma_X = \sigma = \sqrt{\text{Var}(X)}$$

Properties of Variances and Standard Deviations

1. For any constant c, $\text{Var}(c) = 0$.
2. For any constants a and b, $\text{Var}(aX + b) = a^2\text{Var}(X)$.
3. For any constants a and b, $\text{sd}(aX + b) = |a|\sqrt{\text{Var}(X)}$.
4. For any X, $\text{Var}(X) \geq 0$ and $\text{sd}(X) \geq 0$.

Problem 7. $f_X(x) = 2x$ for $0 \leq x \leq 1$ and $f_X(x) = 0$ otherwise. Calculate $\text{Var}(X)$.

Problem 8. Calculate the variance and standard deviation of your earnings upon graduation as described in Problem 6.

4.2.4 Covariance and Correlation

If X and Y are random variables (discrete or continuous), $\mu_X = \mathbb{E}[X]$, and $\mu_Y = \mathbb{E}[Y]$, then their **covariance** is:

$$\text{Cov}(X,Y) = \sigma_{XY} = \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right]$$

Exercise: It is often useful to express the covariance as $\text{Cov}(X,Y) = \mathbb{E}[XY] - \mu_X\mu_Y$. Derive this result from the definition of covariance.

The **correlation coefficient** between X and Y is:

$$\text{Corr}(X,Y) = \rho_{XY} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var(X)}\text{Var}(Y)}}$$

The correlation coefficient is effectively a normalization of covariance. This is closely related to the underlying algebraic formulation of the coefficient recovered from a linear regression of Y on X (or vice versa). Whereas the covariance can take on any numerical value, the value of the correlation coefficient always lies between -1 and 1.

Properties of Covariances and Correlations

1. If X and Y are independent, then $\text{Cov}(X,Y) = 0$. This implies $\text{Corr}(X,Y) = 0$. However, the converse is not true: zero covariance (or a correlation coefficient of zero) does not imply independence.
2. For any constants a_1, a_2, b_1, and b_2, $\text{Cov}(a_1X + b_1, a_2Y + b_2) = a_1a_2\text{Cov}(X,Y)$.
3. $\text{Cov}(X,X) = \text{Var}(X)$.
4. $\text{Cov}(X,Y) = \text{Cov}(Y,X)$.
5. $\text{Cov}(X+Y, W+V) = \text{Cov}(X,W) + \text{Cov}(X,V) + \text{Cov}(Y,W) + \text{Cov}(Y,V)$.
6. The Cauchy-Schwartz Inequality implies: $|\text{Cov}(X,Y)|^2 \leq \text{Var}(X)\text{Var}(Y)$.
7. $-1 \leq \text{Corr}(X,Y) \leq 1$. *Note: This is implied by Property 6.*
8. $\text{Corr}(X,Y)$ does not depend on the units of X and Y.

Problem 9. Consider two independent rolls of a fair die and let X and Y denote the values of the first and second rolls. What is $\mathrm{Cov}(X+Y, X-Y)$?

Are the random variables $X+Y$ and $X-Y$ independent? If not, provide an example showing otherwise. *Hint: Use the definition of independence for discrete random variables and try the realizations $X+Y=12$ and $X-Y=-6$.*

This example confirms covariance property 1.

Problem 10. Let X and Y be discrete random variables with the joint probability mass function

$$f_{X,Y}(x,y) = \frac{1}{4}, \quad (x,y) \in \{(0,0),(1,1),(1,-1),(2,0)\}$$

and zero otherwise. Find the marginal probability mass function, expectation, and variance of X.

Find the marginal probability mass function, expectation, and variance of Y.

Compute $\mathbb{E}[XY]$ and use this to find the covariance of X and Y.

Are X and Y independent?

4.2.5 Variances of Sums of Random Variables

1. For scalar random variables X and Y and any constants a and b, $\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y)$.

2. If X and Y are uncorrelated—that is, if $\text{Cov}(X, Y) = 0$—then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

and

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y).$$

3. For any constants $\{a_1, a_2, \cdots, a_n\}$ and independent random variables $\{X_1, X_2, \cdots, X_n\}$,

$$\text{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i).$$

4. When $a_i = 1\ \forall i$, Property 3 reduces to

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i).$$

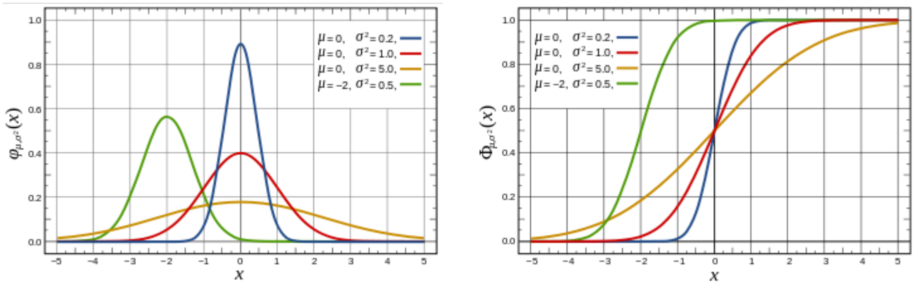
Problem 11. Consider the data points (1, 10), (2, 16), (3, 15), (4, 30), and (5, 14), where each data point $(\cdot, \cdot)$ represents a realization of the random variables (X, Y). Calculate $\text{Cov}(X, Y)$ and $\text{Corr}(X, Y)$ assuming each outcome occurs with equal probability.

4.3 Some Useful Distributions

4.3.1 Normal Distribution

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \quad , \quad -\infty < x < \infty$$



(a) Normal pdf

(b) Normal cdf

Source: "Normal Distribution PDF" by Inductiveload. Licensed under Public domain via Wikimedia Commons.

Properties of the Normal Distribution

1. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.
2. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.
3. If $\{Y_1, Y_2, \cdots, Y_N\}$ are independent RVs and $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ $\forall i$, then $\bar{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$.

Standard Normal Distribution

This is a special case of the normal distribution where $\mu = 0$ and $\sigma^2 = 1$:

$$Z \sim \mathcal{N}(0, 1)$$

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-(z)^2}{2}\right] \quad , \quad -\infty < z < \infty$$

Lognormal Distribution

If X is a positive random variable and $Y = \log(X)$ has a normal distribution, X has a lognormal distribution. This distribution is sometimes used to model non-negative economic variables, such as income and market entry costs (i.e., fixed costs).

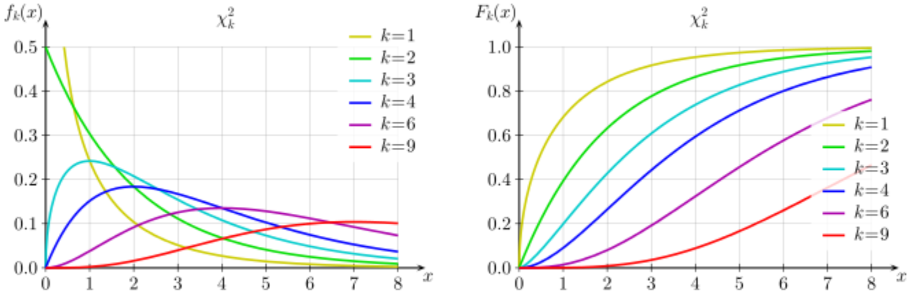
4.3.2 Chi-Squared Distribution

The chi-squared distribution is obtained directly from independent, standard normal random variables. Let $Z_1, Z_2, \cdots, Z_k$ be k independent random variables, each distributed as standard normal. Define a new variable X as the sum of the squares of the Z_i:

$$X = \sum_{i=1}^{k} Z_i^2$$

We say that X has a **chi-squared distribution** with k degrees of freedom: $X \sim \chi_k^2$.

Example: Wald test statistic



(a) Chi-squared pdf

(b) Chi-squared cdf

Source:"Chi-square pdf and cdf" by Geek3 - Own work. Licensed under Creative Commons Attribution 3.0 via Wikimedia Commons.

4.3.3 t Distribution

The t distribution is obtained from a standard normal and a chi-square random variable. Let $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi_k^2$. Furthermore, assume that X and Z are independent. Define a new random variable T:

$$T = \frac{Z}{\sqrt{X/k}}$$

This variable T is distributed according to the t distribution with k degrees of freedom: $T \sim t_k$.

Example: Historically, the t distribution has been used for statistical inference with small sample sizes. This is becoming less common in modern empirical analysis due to larger datasets, but it still comes up on occasion.

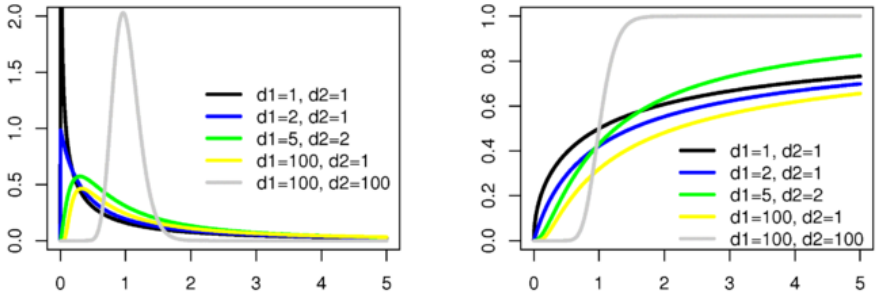
4.3.4 F Distribution

The F distribution is obtained from two chi-square random variables. Let $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$. Furthermore, assume that X_1 and X_2 are independent. Define a new random variable F:

$$F = \frac{X_1/k_1}{X_2/k_2}.$$

We say that F has an F distribution with $(k1, k2)$ degrees of freedom: $F \sim F_{k_1,k_2}$. k_1 is often called the degrees of freedom of the numerator. k_2 is often called the degrees of freedom of the denominator.

Example: The F-statistic is used in econometrics for testing instruments.



(a) F-distribution pdf

(b) F-distribution cdf

4.4 Basic Sampling Theory

In your statistics courses, you will try to learn something about a **population** of subjects (which could be individuals, firms, cities, states, etc.) by analyzing a **sample** of that population.

Let's call the population parameter whose value we want to know θ. Mathematical statistics is largely about designing an estimator for θ.

Estimator: a *rule* (think formula or function) that assigns each possible outcome of the sample a particular value (or distribution of values). Call the estimator W.

Estimate: an actual value generated using the estimator based on a particular sample. We call the estimate $\hat{W}$.

4.4.1 Three Praiseworthy Properties of Estimators

1. Unbiasedness: $\mathbb{E}[W] = \theta$.
2. Efficiency: An estimator W_1 is efficient relative to W_2 if $\text{Var}(W_1) < \text{Var}(W_2)$.
3. Consistency: An estimator W is consistent if $\lim_{n\to\infty} \Pr(|W_n - \theta| \geq \epsilon) = 0$ as the sample size (n) increases to infinity.

Problem 12. Suppose we want to characterize the distribution of heights of women in the United States. Assume the height of women in the United States is normally distributed: $X \sim \mathcal{N}(\mu, \sigma^2)$. If we choose appropriate estimators and collect some data, we can construct estimates of the parameters (μ, σ^2) of this distribution, conditional on our data. Suppose you hired a research assistant to randomly sample heights, and now you have the following data: $\{X_1, X_2, \cdots, X_n\} = \{60, 62, 64, 66, 68\}$, all in inches.

First, consider the average height. One **estimator** for μ is the sample mean: $W_\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$. What is our **estimate** $\hat{W}_\mu$?

Now, consider the variance of heights. One **estimator** for the population variance, σ^2, is the biased sample variance: $W_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{W}_\mu)^2$. What is our **estimate** $\hat{W}_{\sigma^2}$?

Each time we take a sample from the population, the sample will be different. Consequently, our estimate will be different—our estimate is **conditional** on the sample.

Problem 13. Let Y_1, Y_2, Y_3, Y_4 be independent, identically distributed random variables from a population with mean μ and variance σ^2. Let $\bar{Y} = (Y_1 + Y_2 + Y_3 + Y_4)/4$ denote the average of these four random variables.[13]

1. What are the expected value and variance of $\bar{Y}$ in terms of μ and σ^2?

2. Now, consider a different estimator of μ: $W = \frac{Y_1}{8} + \frac{Y_2}{8} + \frac{Y_3}{4} + \frac{Y_4}{2}$. Show that W is also an unbiased estimator of μ. Find the variance of W.

3. Based on your answers to parts (i) and (ii), which estimator of μ do your prefer, $\bar{Y}$ or W? Why?

4. Now, consider a different estimator of μ: $W_a = a_1 Y_1 + a_2 Y_2 + a_3 Y_3 + a_4 Y_4$. What condition is needed on the a_i for W_a to be an unbiased estimator of μ?

[13] Source: Wooldridge 2000, Appendix C, Problem C1.

4.4.2 The Law of Large Numbers (LLN)

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with mean μ. Furthermore, let $\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$. Then

$$\lim_{n \to \infty} \Pr(|\bar{X} - \mu| \geq \epsilon) = 0.$$

In other words, as $n \to \infty$, the mean of a sample of i.i.d. random variables equals the "true" (population) mean. We can also say that the **probability limit** of $\bar{X}$ is μ: $\text{plim}(\bar{X}) = \mu$.

4.4.3 The Central Limit Theorem (CLT)

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with mean μ and variance σ^2. Furthermore, let $\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$. Then, as $n \to \infty$,

$$\sqrt{(n)} \left(\left(\frac{1}{n} \sum_{i=1}^{n} \mathrm{X}_i \right) - \mu \right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

In words, as $n \to \infty$, the mean of a sample is approximately normally distributed *regardless of the underlying distribution of the sample*. We call this **asymptotic normality**.

5 Further Topics: Optimization

Sources for this section can be found in Chiang and Wainwright, chapters 21.1-21.5; Dixit's Optimization in Economic Theory, chapter 3; Simon and Blume, chapters 18.3-18.6; and Martin Osborne's website.

5.1 Checking Quasiconcavity and Quasiconvexity

We said in the main text that we check for quasiconcavity and quasiconvexity by inspection. This, however, may be too difficult to do. Fortunately, if our function is differentiable (twice differentiable in one case), then we can apply the following propositions.[14]

Firstly,

the differentiable function f of n variables defined on a convex set S is quasiconcave on S if and only if

$$u \in S, v \in S, \text{and } f(u) \geq f(v) \implies \sum_{i=1}^{n} f_i'(v) \cdot (u_i - v_i) \geq 0$$

and is quasiconvex on S if and only if

$$u \in S, v \in S, \text{and } f(u) \leq f(v) \implies \sum_{i=1}^{n} f_i'(v) \cdot (u_i - v_i) \leq 0.$$

Secondly,

[14] These proposition are outlined in Osborne's online textbook, chapter 3.4. See the cites therein for proofs.

let f be a twice-differentiable function of n variables. For $r = 1, \ldots, n$, the rth order bordered Hessian of f at the point x is the matrix

$$\begin{bmatrix} 0 & f'_1 & f'_2 & \cdots & f'_r \\ f'_1 & f''_{11} & f''_{12} & \cdots & f''_{1r} \\ f'_2 & f''_{21} & f''_{22} & \cdots & f''_{2r} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ f'_r & f''_{r1} & f''_{r2} & \cdots & f''_{rr} \end{bmatrix}$$

Further, let f be a function of n variables with continuous partial derivatives and cross partial derivatives in an open convex set S and let D_r be the determinant of its rth order bordered Hessian.

- If f is quasiconcave then $D_1(x) \leq 0$, $D_2(x) \geq 0, \ldots, D_n(x) \leq 0$ if n is odd and $D_n(x) \geq 0$ if n is even, for all x in S. (Note that the first condition is automatically satisfied.)
- If f is quasiconvex then $D_k(x) \leq 0$ for $k = 1, ..., n$ for all x in S. (Note that the first condition is automatically satisfied.)
- If $D_1(x) < 0$, $D_2(x) > 0, \ldots, D_n(x) < 0$ if n is odd and $D_n(x) > 0$ if n is even for all x in S then f is quasiconcave.
- If $D_k(x) < 0$ for $k = 1, \ldots, n$ for all x in S then f is quasiconvex.

5.2 Motivation for Inequality Constraints

Many problems in economics have either (1) non-binding constraints, or (2) the possibility of corner solutions.

Examples of non-binding constraints include:

- Non-negativity constraints on production inputs, e.g.:

$$\min_{x_1,x_2} C = (x_1 - 4)^2 + (x_2 - 4)^2$$

s.t.

$$\begin{aligned} 2x_1 + 3x_2 &\geq 6 \\ -3x_1 - 2x_2 &\geq -12 \\ x_1, x_2 &\geq 0 \end{aligned}$$

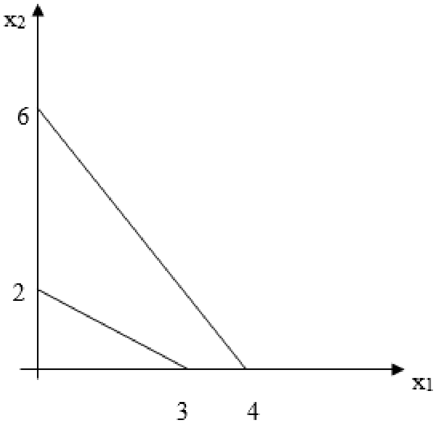


Figure 8: The solution to the problem is confined to be in the area sectioned-off by the equations; it does not necessarily have to lie on a constraint.

- Non-negativity constraints on firm profits.

Examples of corner solutions include:

- Consumer demand for a subset of available goods;
- Production using a subset of available goods.

In a lot of cases, the best way to deal with non-binding constraints is to solve the problem ignoring the constraints and ex-post check that the solutions satisfy these constraints. If the solutions do not satisfy these constraints, we can

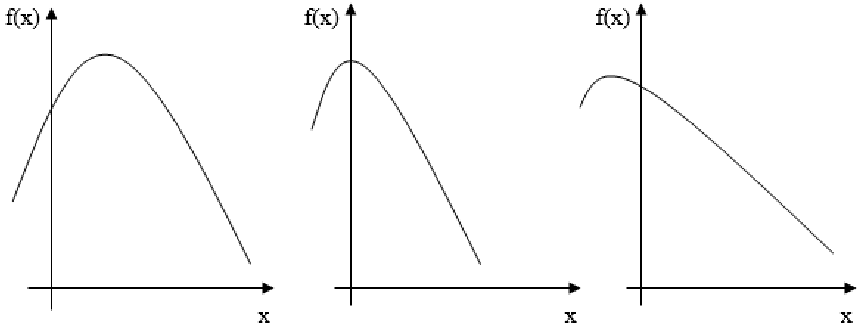


Figure 1. Figure 2. Figure 3.

Figure 9: If we are constrained to pick a positive value of x, we may be stuck with a corner solution.

use Kuhn-Tucker conditions.

5.3 The Kuhn-Tucker Conditions

The Kuhn-Tucker conditions are stated in the following theorem:

Suppose x is an n-dimensional vector, c an m-dimensional vector of scalars (with c_i denoting the ith entry), F a function taking values, G a function taking m-dimensional vector values (with G_i its ith value). Define the Lagrangian,

$$\mathcal{L} = F(x) - \sum_i \lambda_i [G_i(x) - c_i],$$

where λ_i the Lagrangian multiplier on the ith constraint. Suppose x^* maximizes $F(x)$ subject to $G(x) \leq c$ and $x \geq 0$. Then, there is a value of λ such that

$$\frac{\partial \mathcal{L}}{\partial x}(x^*) \leq 0 \text{ and } x^* \geq 0, \text{ with complementary slackness,}$$

and $\forall i$

$$\frac{\partial \mathcal{L}}{\partial \lambda_i}(x^*) \geq 0 \text{ and } \lambda_i \geq 0, \text{ with complementary slackness,}$$

where complementary slackness means that one of the two conditions is binding.

Notice the wording of the theorem; the Kuhn-Tucker conditions are only necessary for a solution to a maximization problem. They are not sufficient. That is, only values that satisfy the Kuhn-Tucker conditions can be solutions to a given maximization problem, but not all values that satisfy the Kuhn-Tucker conditions are solutions to the maximization problem. The theorem is also stated for maximization problems only. To apply this theorem to minimization problems, simply substitute $F(x)$ for $-F(x)$.

Example: Quasi-Linear Preferences (Example 3.1 in Dixit's Optimization in Economic Theory)

Consider an individual with quasi-linear preferences,

$$U(x,y) = y + a\ln(x),$$

where a is a given positive constant. Maximize this consumer's utility subject to their budget constraint and a non-negativity condition of the quantities of goods x and y consumed, with $p > 0$ the price of good x and $q > o$ the price of good y. The consumer has $I > 0$ income to spend. That is,

$$\max_{x,y} y + a\ln(x)$$

s.t.

$$\begin{aligned} x, y &\geq 0 \\ px + qy &\leq I \end{aligned}$$

We now apply the Kuhn-Tucker Theorem. Form the Lagrangian,

$$\mathcal{L} = y + a\ln(x) - \lambda[px + qy - I].$$

The Kuhn-Tucker conditions are,

$$\begin{aligned} \frac{a}{x} - \lambda p \leq 0 &\text{ and } x \geq 0 \\ 1 - \lambda q \leq 0 &\text{ and } y \geq 0 \\ I - px - qy \geq 0 &\text{ and } \lambda \geq 0, \end{aligned}$$

in each case with complementary slackness.

Notice that we have $2^3 = 8$ possible solutions:

$$\begin{aligned} x^* = 0, y^* = 0, \lambda^* = 0 \\ x^* \neq 0, y^* = 0, \lambda^* = 0 \\ x^* \neq 0, y^* \neq 0, \lambda^* \neq 0 \\ x^* = 0, y^* \neq 0, \lambda^* = 0 \\ x^* = 0, y^* \neq 0, \lambda^* \neq 0 \\ x^* = 0, y^* = 0, \lambda^* \neq 0 \\ x^* \neq 0, y^* = 0, \lambda^* \neq 0 \\ x^* \neq 0, y^* \neq 0, \lambda^* = 0 \end{aligned}$$

We can use economic intuition to sort through them.

1. Note that with positive marginal utilities for both x and y, it will always

be beneficial for the consumer to spend more money on either good. Thus, we have $I - px - qy = 0 \implies \lambda \neq 0$. We now sort between,

$$\begin{gathered} x^* \neq 0, y^* \neq 0, \lambda^* \neq 0 \\ x^* = 0, y^* \neq 0, \lambda^* \neq 0 \\ x^* = 0, y^* = 0, \lambda^* \neq 0 \\ x^* \neq 0, y^* = 0, \lambda^* \neq 0 \end{gathered}$$

2. Having both $x^* = y^* = 0 \implies I = 0$, which we have assumed is not true. So we can ignore that case. This leaves us with,

$$\begin{gathered} x^* \neq 0, y^* \neq 0, \lambda^* \neq 0 \\ x^* = 0, y^* \neq 0, \lambda^* \neq 0 \\ x^* \neq 0, y^* = 0, \lambda^* \neq 0 \end{gathered}$$

3. Lastly, see that having $x = 0$ and $y = I/q > 0 \implies \lambda = 1/q$ which in turn implies that $p/q > \infty$. This is impossible. This leaves us with our final candidate solutions:

$$\begin{gathered} x^* \neq 0, y^* \neq 0, \lambda^* \neq 0 \\ x^* \neq 0, y^* = 0, \lambda^* \neq 0 \end{gathered}$$

With some algebra, we find that the optimum choice rule is:

$$\begin{aligned} \text{If } I \leq aq,\ x = I/p \text{ and } y = 0, \\ \text{if } I > aq,\ x = aq/p \text{ and } y = I/q - a. \end{aligned}$$

5.4 Sufficiency of Kuhn-Tucker Conditions

As noted above, satisfying the Kuhn-Tucker conditions does not guarantee that (x^*, λ^*) is a solution to the maximization problem. Consider the following example:

$$\max \; x_1$$

subject to,

$$x_2 - (1 - x_1)^2 \leq 0$$
$$x_1, x_2 \geq 0$$

What is the reason for the above result? Are there conditions which guarantee that (x^*, λ^*) is a solution to the maximization problem? Yes!

The Arrow-Enthoven Sufficiency Theorem

Given the problem:

$$\max \ f(x)$$

subject to

$$g_i(x) \leq c_i$$

$$x \geq 0$$

If

1. The objective function $f(x)$ is differentiable and quasiconcave in the nonnegative orthant;

2. Each constraint function $g_i(x)$ is differentiable and quasiconvex in the nonnegative orthant;

3. The point x^* satisfies the Kuhn-Tucker maximum conditions;

4. Any *one* of the following is satisfied:

   (a) $f_j(x^*) < 0$ for at least one variable x_j;

   (b) $f_j(x^*) > 0$ for some variable x_j that can take on a positive value without violating the constraints;

   (c) The n derivatives $f_j(x^*)$ are not all zero, and the function $f(x)$ is twice differentiable in the neighborhood of x^* (i.e. all the second-order partial derivatives of $f(x)$ exist at x^*);

   (d) The function $f(x)$ is concave;

Then x^* is a global constrained maximum of $f(x)$.

6 Further Topics: Proofs

6.1 Structure of a Proof

Suppose we have a theorem, proposition, or fact we are studying. It typically takes the form:

- A is true.

Or of the form:

- If A is true, then B is true.

A proof is a series of statements that explains, in a logical way, why the theorem, proposition, or fact is true. Each statement in the proof is either:

- An assumption or accepted axiom or theorem, or;
- a conclusion, which follows from the assumptions or previously proven results.

That's it.

This handout covers several common methods of proofs:

- Direct proof
- Proof by contradiction
- Proof by contrapositive
- Proof by induction

It also discusses a few additional topics:

- Proving "if and only if" statements
- Disproving by counterexample

6.2 Direct Proof

Suppose we have a theorem, proposition, or fact of the form:

- A is true.

Or of the form:

- If A is true, then B is true.

A direct proof shows, step by step, why A (or B) is true. A direct proof often employs simple algebra, accepted axioms, and theorems.

Example:

Prove the following: Let m be an even integer and p be any integer. Then mp is an even integer.

Example:

Prove that the least-squares estimator $\hat{\beta} = (X'X)^{-1}X'Y$ is unbiased.

6.3 Proof by Contradiction

Suppose we have a theorem, proposition, or fact of the form:

- If A is true, then B is true.

To construct a proof by contradiction:

- Assume A is true;
- Assume B is false;
- Build upon these assumptions until you reach a contradiction.

The contradiction you reach might be:

- An obvious falsehood, such as 1 = 0.
- A conclusion contradicting one of your assumptions (such as A).

Example:

Prove the following: If p is a prime number bigger than 2, then p is odd.

6.4 Proof by Contrapositive

Note that the following statements are logically equivalent:

- If P, then Q.
- If not Q, then not P.

So, suppose we have a theorem, proposition, or fact of the form:

- If A is true, then B is true.

To construct a proof by contrapositive, you argue that:

- If B is not true, then A is not true.

Proof by contradiction and proof by contrapositive may seem like the same thing, but they are different in a subtle way:

- Proof by Contradiction: Assume A is true and B is not true and show a contradiction.
- Proof by Contrapositive: Assume B is not true and show that A is not true.

Example:

Argue that if it is raining, it must be cloudy.

Example:

Prove that every Giffen good is an inferior good.

6.5 Proof by Induction

To construct a proof by induction:

- State the theorem, proposition, or fact you want to prove.
- Verify the base case. In other words, show that the smallest case is true.
- Assume that the theorem, proposition, or fact is true for some integer k.
- Verify that the theorem, proposition, or fact is true for $(k+1)$.

Proof by induction works because, if the smallest case is true, and the second-smallest case follows from the first, and the third-smallest case follows from the second, then the theorem, proposition, or fact is true.

Example:

Prove that, for any positive integer n, $1+2+3+\cdots+n=\frac{n(n+1)}{2}$.

6.6 Necessary and Sufficient Conditions and Proving If-and-Only-If Statements

Suppose we have two statements, A and B. There are several possible relationships among these statements:

- If B, then A.
- If A, then B.
- A if and only if B.

The first possibility, "If B, then A," implies that A is a necessary condition for B. This means that A must be true for B to be true. In other words, B cannot be true if A is not true.

The second possibility, "If A, then B," implies that A is a sufficient condition for B. This means that B is true if A is true. However, B can still be true even if A is not true. In other words, A is not a necessary condition for B.

The third possibility, "A if and only if B," implies that A is both a necessary and a sufficient condition for B. "A if and only if B" means that both of the first two possibilities hold:

- If B, then A, and
- If A, then B.

To prove an "if and only if" statement, you must show that both of these possibilities are true.

Example:

It is the month of February if and only if there are fewer than 30 days in the month.

Example:

An economic agent will prefer lottery L_1 to another lottery L_2 if and only if the expected utility of L_1 is greater than the expected utility of L_2.