NOTES FOR QUANTITATIVE METHODS:
SOME ADVANCED MATHEMATICS FROM AN ELEMENTARY POINT OF VIEW

MAXWELL B. STINCHCOMBE

Contents

Date: November 25, 2003.
Fall Semester, 2003. Unique #30238.

0. Organizational Stuff

Meetings: Mondays and Wednesdays, 2-3:30 and Wednesdays 8:30-9:30, in BRB 1.120.

Teachers: My office is BRB 2.118, phone number is 475-8515, e-mail address is `maxwell@eco.utexas.ed` office hours Mondays and Wednesdays 10-12. You are very lucky to have Lori Stuntz as the T.A. for this course. Her office is 4.116, e-mail address is `stuntz@eco.utexas.edu`, office hours TBA.

Texts: For the statistical part of the course, we'll use George Casella and Roger Berger's *Statistical Inference*, 2'nd ed. (Duxbury 2002), following it fairly closely. For the optimization and analysis parts of the course, we'll use Sheldon M. Ross's *Applied Probability Models with Optimization Applications* (Dover Publications, 1992) and A. N. Kolmogorov and S. V. Fomin's *Introductory Real Analysis* (Dover Publications, 1970). Throughout, you will be refering to the microeconomics textbook, *Microeconomic Theory*, by Mas-Colell, Whinston, and Green.

Topics: Completeness properties of $\mathbb{R}$ and $\mathbb{R}^\ell$, summability and valuation of streams of utilities, convex analysis and duality; further properties of $\mathbb{R}^\ell$ and related spaces, (including compactness, continuity and measurability of functions on $\mathbb{R}^\ell$, summability of sequences, existence of optima, fixed point theorems, cdf's, other metrics, other metric spaces, the Theorem of the Maximum); Probabilities and expectations (including domains, modes of convergence, convergence theorems, orders of stochastic dominance, conditional expectations and probabilities); Dynamic programming (including properties of sequence spaces and probabilities on them, Bellman and Euler equations, the role of the Theorem of the Maximum, growth models); Statistics (including specific distributions [uniform, gamma, beta, Gaussian, t, F, χ^2, Poisson, negative exponential, Weibull, logistic], estimators and their properties [consistency, Glivenko-Cantelli, different kinds of "best" estimators, Bayesian estimators, MLE estimators, information inequalities, sufficiency, Blackwell-Rao], properties of hypothesis tests [types of errors and their associated distributions, the Neyman-Pearson Lemma]).

1. Some Basics About Numbers and Quantities

Readings: Marinacci's "An Axiomatic Approach to Complete Patience and Time Invariance," *Journal of Economic Theory* **83**, 105-144 (1998). Mas-Colell, Whinston, and Green on support functions and the supporting hyperplane theorem. §1.6 below is for you to read and work on, either by yourself or in a study group.

1.1. **Lengths and measurements.** $\mathbb{N}$ and $\mathbb{Q}$ from elementary school. As models of measurements of quantities, we're done.

1.2. **Why we want more.** Irrationality of easily described lengths, clt and integration.

Sequences in $\mathbb{Q}$, convergence implies settling down, but not the reverse. Subsequences.

Cauchy sequences and $\mathbb{R}$ as the completion of $\mathbb{Q}$.

Implications of completeness: decreasing and increasing bounded sequences have limits, equivalently, every bounded set has a sup and an inf. The idea of completion also shows up in the major limit theorem in statistics (i.e. the CLT).

1.3. **Valuing sequences of rewards.** This section is based on classic analyses as well as the more recent Marinacci's "An Axiomatic Approach to Complete Patience and Time Invariance," *Journal of Economic Theory* **83**, 105-144 (1998). Patience about finite sequences, $(r_1, r_2, \ldots, r_t)$, of rewards seems to be about being indifferent between all time permutations of the sequence. In the dynamic programming models used in game theory and macro, one often achieves infinite sequences of rewards. These may not be entirely believable, but they do a pretty good job of capturing the idea of an indefinite future.

1.3.1. *Classic analyses.* $\liminf_t r_t \leq \limsup_t r_t$, equality for limits.

X^Y notation, e.g.'s 2^3, $\mathbb{R}^\ell$, $\mathbb{R}^{\mathbb{N}}$.

Summability for $x \in \mathbb{R}^{\mathbb{N}}$.

$u : \mathbb{R} \to \mathbb{R}$, $u(x) := (u(x_t)_{t\in\mathbb{N}}) \in \mathbb{R}^{\mathbb{N}}$, u **bounded**. From here on, we'll just use x for the elements in $\mathbb{R}^{\mathbb{N}}$ and try to value them, thinking that they are bounded streams of utilities.

$V_{\liminf}(x) := \liminf_t x_t$ (that is, $\liminf_t u(x_t)$) which always exists (by completeness). The infinite extension of a simple of idea of patience is here — any permutation of the integers fails to change $V_{\liminf}(x)$.

Thinking about a finite sequence of rewards, a useful definition of patience is that any permutation of the reward sequence is indifferent, having the good stuff early is just as desirable as having it late. If $\pi : \mathbb{N} \to \mathbb{N}$ is 1-1 and onto, then x_π denotes the sequence $(x_{\pi(t)})_{t\in\mathbb{N}}$. π is a **permutation** of $\mathbb{N}$.

Theorem 1.3.1. *For any permutation π, $V_{\liminf}(x) = V_{\liminf}(x_\pi)$.*

Other ideas of patience include taking the limit of time averages and discounting with a discount factor close to 1. $V_1(x) := \lim_T \frac{1}{T}\sum_{t=1}^T x_t$ exists iff $V_2(x) := \lim_{\delta\uparrow 1}(1-\delta)\sum_{t=1}^\infty \delta^t x_t$ exists, in which case $V_1(x) = V_2(x)$ (Froebenius and Littlewood).

The sequence of x_t

$$0101010101010101\cdots$$

has $V_{\liminf}(x) = 0 < V_1(x) = V_2(x) = \frac{1}{2}$.

1.3.2. *Marinacci's extension of the classic analyses.* The sequence of x_t

$$\underbrace{11}_{2^1}\underbrace{0000}_{2^2}\underbrace{11111111}_{2^3}\cdots$$

is of the form x_π for a peculiar kind of π. Further $V^{average}_{\liminf}(x) := \liminf_T \frac{1}{T}\sum_{t=1}^T x_t = \frac{1}{3} < V^{average}_{\limsup}(x) := \limsup_T \frac{1}{T}\sum_{t=1}^T x_t = \frac{2}{3}$. An alternate criterion combining the averaging and minimizing is $V_3(x) := \lim_{T\to\infty}\{\inf_{j\geq 1}\frac{1}{T}\sum_{t=1}^T x_{t+j}\}$. Note $V_3(x) = 0 < \frac{1}{3}$.

The π above is "peculiar" because it does not preserve upper densities — $A \subset \mathbb{N}$ has a **natural density** $\delta(A) := \lim_T \frac{1}{T}\#\{t \leq T : t \in A\}$ when the limit exists, otherwise A does not have a density. A permutation preserves x's upper densities if for all $r \in \mathbb{R}$, $\delta(\{t : x_t \geq r\}) = \delta(\{t : x_{\pi(t)} \geq r\})$.

Marinacci defines patience as invariance under permutations that preserve upper densities, and asks for all such π, $V(x) = V(x_\pi)$. He arrives at the Polya criterion —

$$V_{Polya}(x) := \lim_{\epsilon\to 0}\left[\liminf_T \frac{1}{\epsilon T}\sum_{t=(1-\epsilon)T}^{T} x_t\right].$$

$V_{Polya}(x)$ exists for all $x \in \mathbb{R}^{\mathbb{N}}$, and for x such that $V^{average}_{\liminf}(x) = V^{average}_{\limsup}(x)$, $V_{Polya}(x) = V^{average}_{\lim}(x)$.

1.4. **Convex analysis.** $\mathbb{R}^\ell$, vectors. Two important examples of convex sets from microeconomics:

1. Netput vectors in production, intermediate micro $x_2 \leq f(x_1)$, x_1 the input, x_2 the output, becomes $Y = \{(y_1, y_2) : y_1 \leq 0,\ y_2 \leq f(|y_1|)\}$.
2. Given a utility function $u : \mathbb{R}^\ell_+ \to \mathbb{R}$ and a utility level $\overline{u}$, $C_u(\overline{u}) := \{x \in \mathbb{R}^\ell_+ : u(x) \geq \overline{u}\}$ is a upper contour set of u. Decreasing marginal rates of substitution are captured by the assumption that each $C_u(\overline{u})$ is a convex set.

The extreme values of linear functions over convex sets play a crucial role in neoclassical economics.

In the first example above, the profit function is $\pi(p) := \max\{p \cdot y : y \in Y\}$ where $p \gg 0$ is a price vector. In the second example above, the expenditure function is $e(p, \overline{u}) :=$

$\min\{p \cdot x : x \in C_u(\overline{u})\}$, again, $p \gg 0$ a price vector. One of the things we will see is that $D_p\pi = y^*(p)$ where $y^*(p)$ is the solution to $\max\{p \cdot y : y \in Y\}$, and $D_p e(p, \overline{u}) = h(p, \overline{u})$ where $h(p, \overline{u})$ is the solution to $\min\{p \cdot x : x \in C_u(\overline{u})\}$. $y^*(p)$ is the supply/demand function for the firm, $h(p, \overline{u})$ is the Hicksian demand function.

1.4.1. *Convexity of sets.* Dfn **convexity** of sets, e.g.'s hyperplanes $H_p^{\leq}(r) := \{x : p \cdot x \leq r\}$, triangles, squares, the set Y above when $f'' < 0$. A function is **concave** when its subgraph is a convex set. In the production function example, you should recognize decreasing returns to scale (DRTS). When the epigraph is a convex set, the function is convex. There is no such creature as a concave set.

1.4.2. *Three basic results.*

Theorem 1.4.1. *If $\{K_\alpha : \alpha \in A\}$ is a collection of convex sets, then $K := \cap_\alpha K_\alpha$ is convex.*

Proof.

Corollary 1.4.1.1. *If $u : \mathbb{R}^\ell_+ \to \mathbb{R}$ is concave, then for each $\overline{u}$, $C_u(\overline{u}) := \{x \in \mathbb{R}^\ell_+ : u(x) \geq \overline{u}\}$ is a convex set.*

Proof.

In intermediate micro, one starts with a utility function, u, that represents preferences, that is, $x \succsim y$ iff $u(x) \geq u(y)$, and then derives demand behavior, $x(p, m)$ from the solutions to

$$\max\ u(x)\ \text{ subject to }\ p \cdot x \leq m,\ x \in \mathbb{R}^\ell_+.$$

The demand function, $x(p, m)$, is unaffected by monotonic transformations of u, that is, if $f : \mathbb{R} \to \mathbb{R}$ satisfies $[[r > s] \Rightarrow [f(r) > f(s)]]$ and $v(x) := f(u(x))$, then $x(p, m)$ also solves the problem

$$\max\ v(x)\ \text{ subject to }\ p \cdot x \leq m,\ x \in \mathbb{R}^\ell_+.$$

Utility functions do not measure anything. They are no more than a convenient device to represent preferences.

The property that $C_u(\overline{u})$ is a convex set is preserved under monotonic transformations of u, that is, for every $\overline{v}$, $C_v(\overline{v}) := \{x \in \mathbb{R}^\ell_+ : v(x) \geq \overline{v}\}$ is a convex set. This leads to a definition, with K a convex subset of $\mathbb{R}^\ell$, $v : K \to \mathbb{R}$ is **quasi-concave** if for all $\overline{v}$, $\{x \in K : v(x) \geq \overline{v}\}$ is a convex set.

Two sets, E, E' are disjoint if $E \cap E' = \emptyset$. Stronger than disjoint is being at a positive distance. For $A, B \subset \mathbb{R}^\ell$, $A + B := \{a + b : a \in A,\ b \in B\}$, draw some pictures, show that A, B convex $\Rightarrow A + B$ is convex. For $x \in \mathbb{R}^\ell$ and $\epsilon > 0$, $B(x, \epsilon) := \{y : \|x - y\| < \epsilon\}$. Two sets, $E, E' \subset \mathbb{R}^\ell$ are ϵ-separated if $E + B(0, \epsilon)$ and $E' + B(0, \epsilon)$ are disjoint.

Theorem 1.4.2 (Separating Hyperplane)**.** *If K and K' are disjoint convex subsets of $\mathbb{R}^\ell$, then $\exists p \in \mathbb{R}^\ell$, $p \neq 0$, such that $\forall x \in K, x' \in K'$, $p \cdot x \leq p \cdot x'$. If K and K' are also ϵ-separated, then $\exists p \in \mathbb{R}^\ell$, $p \neq 0$, and $\exists \delta > 0$ such that $\forall x \in K, x' \in K'$, $p \cdot x + \delta \leq p \cdot x'$.*

Pictures of what this means, proof will come later. An interesting application uses the idea of a closed set, intuitively, one containing its boundary. In order to get to interesting economics, I am providing an interim definition of closed sets, one that applies only to convex sets. We will return to the idea of closed sets later.

Definition 1.4.3. *The* **convex-closure of** $K \subset \mathbb{R}^\ell$ *is* $\overline{K} := \bigcap\{H_p^{\leq}(r) : K \subset H_p^{\leq}(r)\}$. *A set K is* **convex-closed** *if $K = \overline{K}$.*

The class of convex-closed sets is closed under intersection, that is, if K_α, $\alpha \in A$, is a collection of convex-closed sets, then $\cap_\alpha K_\alpha$ is convex-closed.

Pictures. The SHThm gives us

Lemma 1.4.4. *If K is convex, then for all $\epsilon > 0$, $\overline{K} \subset K + B(0, \epsilon)$.*

1.4.3. *A worked example.* $Y = \{(y_1, y_2) : y_1 \leq 0,\ y_2 \leq \sqrt{|y_1|}\}$, $\Pi_Y(p) := \sup\{p \cdot y : y \in Y\}$, find the input demand function, the supply function, the profit function, show that the profit function is convex, "application" to stability of prices, refer to homework on derivative tests.

Given a convex profit function $\Pi(\cdot)$, $Y_\Pi := \{y : \forall p > 0,\ p \cdot y \leq \Pi(p)\}$, relate to convex-closed sets, do the work in the example to show the basic duality result for profit functions,

$$Y = Y_{\Pi_Y}.$$

This means that I can give you a profit function and I have implicitly specified the technology, or I can give you a technology, and I have implicitly specified the profit function, and these two representations are (loosely) **duals** of each other.

Applications of this idea to expenditure functions (recovering upper contour sets from expenditure functions and vice versa), costs functions (same idea).

1.4.4. *Support functions.* The inf-support function of a set K is $\mu_K^{\inf}(p) := \inf\{p \cdot x : x \in K\}$. [Beware: this is the support function that most people use, not the next one.] The sup-support function of a set K is $\mu_K^{\sup}(p) := \sup\{p \cdot x : x \in K\}$. Note that $\mu_K^{\sup}(p) = -\mu_K^{\inf}(-p)$, so these are essentially the same function.

Conventions with $\pm\infty$ and $0 < \alpha < 1$ in the definition of concave and convex functions.

Theorem 1.4.5. *An inf-support function is concave, a sup-support function is convex.*

Corollary 1.4.5.1. *An expenditure function is concave, and a profit function is convex.*

Behavioral implications.

The general duality theorem relating closed convex sets to their support functions is

Theorem 1.4.6. *If K is convex-closed, then $K = \{x : \forall p,\ p \cdot x \leq \mu_K^{\sup}(p)\}$.*

This is what we did with the profit function above.

Another important duality result is that the derivative of the support function is the solution to the optimization problem. Suppose that K is convex and that the problem $\max\{p^\circ \cdot x : x \in K\}$ has a unique solution, x°. It can be shown that $\mu_K^{\sup}(\cdot)$ is differentiable at p°. Assuming (but not proving) that differentiability, we will argue that $D_p\mu_K^{\sup}(p^\circ) = x^\circ$. To see why, look at the function $\xi(p) := \mu_K^{\sup}(p) - p \cdot x^\circ$, $\xi(p) \geq 0$ and $\xi(p^\circ) = 0$. This implies that $D_p\xi(p^\circ) = 0$, and $D_p\xi(p^\circ) = D_p\mu_K^{\sup}(p^\circ) - x^\circ$.

1.5. **Problems.**

Problem 1.1. *Show that the sequence of x_t*

$$0101010101010101\ \cdots$$

has $V_{\liminf}(x) = 0 < V_1(x) = V_2(x) = \frac{1}{2}$.

Problem 1.2. *Show that the sequence of x_t*

$$\underbrace{11}_{2^1}\underbrace{0000}_{2^2}\underbrace{11111111}_{2^3}\cdots$$

is a permutation of the sequence

$$0101010101010101\ \cdots.$$

Further show that $V^{average}_{\liminf}(x) = \frac{1}{3} < V^{average}_{\limsup}(x) = \frac{2}{3}$, and that $V_3(x) = 0$.

Problem 1.3. *Let $Y = \{(y_1, y_2) : y_1 \leq 0,\ y_2 \leq \log(1 - y_1)\}$, find the supply and demand function, the profit function Π_Y, and explicitly show that $Y = Y_{\Pi_Y}$.*

Problem 1.4. *Prove Lemma* 1.4.4.

Problem 1.5. *Find the function $\mu_K^{\sup}(p)$ when*

1. $K = B(0, \epsilon)$,
2. $K = B(x, \epsilon)$,
3. $K = H_p^{\leq}(r)$,
4. $K = \{x, y\}$, $x, y \in \mathbb{R}^\ell$, $x \neq y$,
5. $K = \{\alpha x + (1 - \alpha)y : 0 \leq \alpha \leq 1\}$, $x, y \in \mathbb{R}^\ell$, $x \neq y$,
6. $K = \mathbb{R}^\ell_+$.

Problem 1.6. *Show that for all K, $\mu_K^{\sup}(p) = \mu_{\overline{K}}^{\sup}(p)$.*

Problem 1.7 (Cauchy-Schwarz inequality and dot products)**.** *Consider vectors* $x = (x_k)_{k=1}^n$, $y = (y_k)_{k=1}^n$ *and* $z = (z_k)_{k=1}^n$ *in* $\mathbb{R}^n$*. The (dot) product is defined by* $x \cdot y := \sum_k x_k y_k$*, sometimes written as* xy*. Following the logic of Pythagoras's theorem, the length of a vector is* $\|x\| := \sqrt{\sum_k x_k^2} = \sqrt{x \cdot x}$*.*

1. *Show that* $xy = \|x\|\|y\| \cos\theta$ *where* θ *is the angle determined by* x *and* y*.* [*Hint: Drop a perpendicular from* x *to the line spanned by* y *to get the point* ty *for some* $t \in \mathbb{R}$*. Use Pythagoras's theorem and the definition of the cosine to relate* $\|x\|$*,* $\|ty\|$ *and* $\|x - ty\|$*. Rearrange.*]
2. *Sketch at the following sets for at least three values of* r*:*
   (a) $H_y^{\leq}(r) = \{x \in \mathbb{R}^2 : x \cdot y \leq r\}$*,* $y = (1, 2)^T$*,*
   (b) $H_y^{\leq}(r) = \{x \in \mathbb{R}^2 : x \cdot y \leq r\}$*,* $y = (3, 1)^T$*, and*
   (c) $H_y^{\leq}(r) = \{x \in \mathbb{R}^2 : x \cdot y \geq r\}$*,* $y = (-1, 3)^T$*.*
3. *A further rearrangement of* $xy = \|x\|\|y\| \cos\theta$ *gives the Cauchy-Schwartz inequality,* $(\sum_k x_k y_k)^2 \leq (\sum_k x_k^2)(\sum_k y_k^2)$*. Under what conditions is the inequality satisfied as an equality?*
4. *Define* $d(x, y) = \|x - y\|$ *so that* $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$*. Show that* $(\mathbb{R}^n, d)$ *is a metric space.* [*That is, show that* $d(x, y) = d(y, x)$*,* $d(x, y) = 0$ *iff* $x = y$*, and* $d(x, y) + d(y, z) \geq d(x, z)$*. The hardest part is the last inequality, known as the triangle inequality.*]
5. *Define* $\rho(x, y) = \sum_k |x_k - y_k|$ *so that* $\rho : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$*. Show that* $(\mathbb{R}^n, \rho)$ *is a metric space.*

1.6. **Self-guided tour to differentiability and concavity.** This section develops the negative semi-definiteness of the matrix of second derivatives as being equivalent to the concavity of a twice continuously differentiable function. It also develops the determinant test for negative semi-definiteness. Before reading this, you should know the equivalent of the math camp review of matrix multiplication and determinants.

You are responsible for handing in the problems scattered throughout this section by the middle of the semester. I would recommend that you do it before that.

1.6.1. *The two results.* Before giving the results, we need some terminology.

A function $f : C \to \mathbb{R}$ is **strictly concave** if $\forall x, y \in C$, $x \neq y$, and all $\alpha \in (0,1)$, $f(\alpha x + (1-\alpha)y) > \alpha f(x) + (1-\alpha)f(y)$.

A symmetric matrix $n \times n$ matrix $\mathbf{A} = (a_{ij})_{i,j=1,\dots,n}$ is **negative semi-definite** if for all vectors $z \in \mathbb{R}^n$, $z^T \mathbf{A} z \leq 0$, it is **negative definite** if for all $z \neq 0$, $z^T \mathbf{A} z < 0$.

Theorem 1.6.1. *A twice continuously differentiable $f : \mathbb{R}^n \to \mathbb{R}$ defined on an open, convex set C is concave (respectively strictly concave) iff for all $x^\circ \in C$ $D_x^2 f(x^\circ)$ is negative semi-definite (respectively negative definite).*

The **principal sub-matrices** of a symmetric $n \times n$ matrix $\mathbf{A} = (a_{ij})_{i,j=1,\dots,n}$ are the $m \times m$ matrices $(a_{ij})_{i,j=1,\dots,m}$, $m \leq n$. Thus, the 3 principal sub-matrices of the 3×3 matrix

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 4 & \sqrt{3} \\ 0 & \sqrt{3} & 6 \end{bmatrix}$$

are

$$\begin{bmatrix} 3 \end{bmatrix}, \quad \begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix}, \text{ and } \quad \begin{bmatrix} 3 & 0 & 0 \\ 0 & 4 & \sqrt{3} \\ 0 & \sqrt{3} & 6 \end{bmatrix}.$$

Theorem 1.6.2. *A matrix $\mathbf{A}$ is negative semi-definite (respectively negative definite) iff the sign of m'th principal sub-matrix is either 0 or -1^m (respectively, the sign of the m'th principal sub-matrix is -1^m). It is positive semi-definite (respectively positive definite) if you replace "-1^m" with "$+1^m$" throughout.*

In the following two problems, use Theorem 1.6.1 and 1.6.2.

Problem 1.8. *The function $f : \mathbb{R}^2_+ \to \mathbb{R}$ defined by $f(x,y) = x^\alpha y^\beta$, $\alpha, \beta > 0$, is strictly concave on $\mathbb{R}^2_{++}$ if $\alpha + \beta < 1$, and is concave on $\mathbb{R}^2_{++}$ if $\alpha + \beta = 1$.*

Problem 1.9. *The function $f : \mathbb{R}^2_+ \to \mathbb{R}$ defined by $f(x,y) = (x^p + y^p)^{1/p}$ is convex on $\mathbb{R}^2_{++}$ if $p \geq 1$ and is concave if $p \leq 1$.*

1.6.2. *The one dimensional case, $f : \mathbb{R}^1 \to \mathbb{R}$.*

Problem 1.10. *Suppose that $f : (a, b) \to \mathbb{R}$ is twice continuously differentiable.* [*Read the third part of this before starting the first two.*]

1. *Show that if $f''(x) \leq 0$ for all $x \in (a, b)$, then f is concave.* [*Hint: We know that f' is non-increasing. Pick x, y with $a < x < y < b$ and pick $\alpha \in (0, 1)$, define $z = \alpha x + (1 - \alpha)y$. Note that $(z - x) = (1 - \alpha)(y - x)$ and $(y - z) = \alpha(y - x)$. Show*

$$f(z) - f(x) = \int_x^z f'(t)\, dt \geq f'(z)(z - x) = f'(z)(1 - \alpha)(y - x),$$

$$f(y) - f(z) = \int_z^y f'(t)\, dt \leq f'(z)(y - z) = f'(z)\alpha(y - x).$$

   Therefore,

$$f(z) \geq f(x) + f'(z)(1 - \alpha)(y - x),\ f(z) \geq f(y) - f'(z)\alpha(y - x).$$

   Multiply the lhs by α, the rhs by $(1 - \alpha)$, and ... .]
2. *Show that if f is concave, then $f''(x) \leq 0$ for all $x \in (a, b)$.* [*If not, then $f''(x^\circ) > 0$ for some $x^\circ \in (a, b)$ which implies that f'' is strictly positive on some interval $(a', b') \subset (a, b)$. Reverse the above argument.*]
3. *Repeat the previous two problems for strict concavity, changing whatever needs to be changed.*

1.6.3. *The multi-dimensional case, $f : \mathbb{R}^n \to \mathbb{R}$.*

Problem 1.11. *Suppose that $f : C \to \mathbb{R}$ is twice continuously differentiable, C an open convex subset of $\mathbb{R}^n$.*

For each $y, z \in \mathbb{R}^n$, define $g_{y,z}(\lambda) = f(y + \lambda z)$ for those λ in the interval $\{\lambda : y + \lambda z \in C\}$.

1. *Show that f is (strictly) concave iff each $g_{y,z}$ is (strictly) concave.*
2. *Show that $g''(\lambda) = z^T D_x^2 f(x^\circ) z$ where $x^\circ = y + \lambda z$.*
3. *Conclude that f is (strictly) concave iff for all $x^\circ \in C$, $D^2 f(x^\circ)$ is negative semi-definite (negative definite).*

1.6.4. *A fair amount of matrix algebra background.* The previous has demonstrated that we sometimes want to know conditions on $n \times n$ symmetric matrices $\mathbf{A}$ such that $z^T \mathbf{A} z \leq 0$ for all z, or $z^T \mathbf{A} z < 0$ for all $z \neq 0$. We are trying to prove that a $\mathbf{A}$ is negative semi-definite (respectively negative definite) iff the sign of m'th principal sub-matrix is either 0 or -1^m (respectively, the sign of the m'th principal sub-matrix is -1^m). This will take a longish detour through eigenvalues and eigenvectors. The detour is useful for the study of linear regression too, so this section is also background for next semester's econometrics course.

Throughout, all matrices have only real number entries.

$|\mathbf{A}|$ denotes the determinant of the square $\mathbf{A}$. Recall that $\mathbf{A}$ is invertible, as a linear mapping, iff $|\mathbf{A}| \neq 0$. (If these statements do not make sense to you, you missed the math camp and need to do some review.)

Problem 1.12. *Remember, or look up, how to find determinants for* 2×2 *and* 3×3 *matrices.*

A vector $x \neq 0$ is an **eigenvector**[1] and the number $\lambda \neq 0$ is an **eigenvalue** for $\mathbf{A}$ if $\mathbf{A}x = \lambda x$. Note that $\mathbf{A}x = \lambda x$ iff $\mathbf{A}(rx) = \lambda(rx)$ for all $r \neq 0$. Therefore, we can, and do, normalize eigenvectors by $\|x\| = 1$, which corresponds to setting $r = 1/\|x\|$. There is still some ambiguity, since we could just as well set $r = -1/\|x\|$.

In general, one might need to consider λ's and x's that are imaginary numbers, that is $\lambda = a + bi$ with $i = \sqrt{-1}$. This means that x will need to be imaginary too. To see why, read on.

Lemma 1.6.3. $\mathbf{A} = \lambda x$, $x \neq 0$, *iff* $(\mathbf{A} - \lambda\mathbf{I})x = 0$ *iff* $|\mathbf{A} - \lambda\mathbf{I}| = 0$.

Proof: You should know why this is true. If not, you need some more review. ∎

Define $g(\lambda) = |\mathbf{A} - \lambda\mathbf{I}|$ so that g is an n'th degree polynomial in λ. The fundamental theorem of algebra tells us that any n'th degree polynomial has n roots, counting multiplicities, in the complex plane. To be a bit more concrete, this means that there are complex numbers λ_i, $i = 1, \ldots, n$ such that

$$g(y) = (\lambda_1 - y)(\lambda_2 - y)\cdots(\lambda_n - y).$$

The "counting multiplicities" phrase means that the λ_i need not be distinct.

Problem 1.13. *Using the quadratic formula, show that if* $\mathbf{A}$ *is a symmetric* 2×2 *matrix, then both of the eigenvalues of* $\mathbf{A}$ *are real numbers. Give a* 2×2 *non-symmetric matrix with real entries having two imaginary eigenvalues.* [*This can be done with a matrix having only* 0*'s and* 1*'s as entries.*]

The conclusion about real eigenvalues in the previous problem is true for general $n \times n$ matrices, and we turn to this result.

From your trigonometry class (or from someplace else), $(a+bi)(c+di) = (ac-bd)+(ad+bd)i$ defines multiplication of complex numbers, and $(a + bi)^* := a - bi$ defines the complex conjugate of the number $(a + bi)$. Note that $rs = sr$ and $r = r^*$ iff r is a real number for complex r, s. By direct calculation, $(rs)^* = r^*s^*$ for any pair of complex numbers r, s. Complex vectors are vectors with complex numbers as their entries. Their dot product is defined in the usual way, $x \cdot y := \sum_i x_i y_i$. Notationally, $x \cdot y$ may be written $x^T y$. The next proof uses

[1] "Eigen" is a german word meaning "own."

Problem 1.14. *If r is a complex number, then $rr^* = 0$ iff $r = 0$. If x is a complex vector, then $x^T x^* = 0$ iff $x = 0$.*

Lemma 1.6.4. *Every eigenvalue of a symmetric $\mathbf{A}$ is real, and distinct eigenvectors are real, and orthogonal to each other.*

Proof: The eigenvalue part: Suppose that λ is an eigenvalue and x an associated eigenvector so that

$$\mathbf{A}x = \lambda x. \tag{1}$$

Taking the complex conjugate of both sides,

$$\mathbf{A}x^* = \lambda^* x^* \tag{2}$$

because $\mathbf{A}$ has only real entries.

$$[\mathbf{A}x = \lambda x] \Rightarrow [(x^*)^T \mathbf{A}x = (x^*)^T \lambda x = \lambda x^T x^*],$$

$$[\mathbf{A}x^* = \lambda^* x^*] \Rightarrow [x^T \mathbf{A}x^* = x^T \lambda^* x^* = \lambda^* x^T x^*].$$

Subtracting,

$$(x^*)^T \mathbf{A}x - x^T \mathbf{A}x^* = (\lambda - \lambda^*) x^T x^*.$$

Since the matrix $\mathbf{A}$ is symmetric,

$$(x^*)^T \mathbf{A}x - x^T \mathbf{A}x^* = 0.$$

Since $x \neq 0$, $x^T x^* \neq 0$. Therefore,

$$[(\lambda - \lambda^*) x^T x^* = 0] \Rightarrow [(\lambda - \lambda^*) = 0],$$

which can only happen if λ is a real number.

The eigenvector part: From the previous part, all eigenvalues are real. Since $\mathbf{A}$ is real, this implies that all eigenvectors are also real.

Let $\lambda_i \neq \lambda_j$ be distinct eigenvalues and x_i, x_j their associated eigenvectors so that

$$\mathbf{A}x_i = \lambda_i x_i, \quad \mathbf{A}x_j = \lambda_j x_j.$$

Pre-multiplying by the appropriate vectors,

$$x_j^T \mathbf{A}x_i = \lambda_i x_j^T x_i, \quad x_i^T \mathbf{A}x_j = \lambda_j x_i^T x_j.$$

We know that $x_i^T x_j = x_j^T x_i$ (by properties of dot products). Because $\mathbf{A}$ is symmetric,

$$x_j^T \mathbf{A}x_i = x_i^T \mathbf{A}x_j.$$

Combining,

$$(\lambda_i - \lambda_j) x_j^T x_i = 0.$$

Since $(\lambda_i - \lambda_j) \neq 0$, we conclude that $x_i \cdot x_j = 0$, the orthogonality we were looking for. ∎

The following uses basic linear algebra definitions.

Problem 1.15. *If the $n \times n$ $\mathbf{A}$ has n distinct eigenvalues, then its eigenvectors form an orthonormal basis for $\mathbb{R}^n$.*

A careful proof shows that if $\mathbf{A}$ has an eigenvalue λ_i with multiplicity $k \geq 2$, then we can pick k orthogonal eigenvectors spanning the k-dimensional set of all x such that $\mathbf{A}x = \lambda_i x$. There will be infinitely many different ways of selecting such an orthogonal set. You either accept this on faith or go review a good matrix algebra textbook.

Problem 1.16. *Find eigenvalues and eigenvectors for*

$$\begin{bmatrix} 4 & \sqrt{3} \\ \sqrt{3} & 6 \end{bmatrix} \quad \textit{and} \quad \begin{bmatrix} 3 & 0 & 0 \\ 0 & 4 & \sqrt{3} \\ 0 & \sqrt{3} & 6 \end{bmatrix}.$$

Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $\mathbf{A}$ (repeating any multiplicities), and let $u_1, \ldots, u_n$ be a corresponding set of orthonormal eigenvectors. Let $\mathbf{Q} = (u_1, \ldots, u_n)$ be the matrix with the eigenvectors as columns. Note that $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ so that $\mathbf{Q}^{-1} = \mathbf{Q}^T$. A matrix with its transpose being its inverse is an **orthogonal matrix**. Let Λ be the $n \times n$ matrix with $\Lambda_{ii} = \lambda_i$ and with 0's in the off-diagonal.

Problem 1.17. *Show that* $\mathbf{Q}^T\mathbf{A}\mathbf{Q} = \Lambda$*, equivalently,* $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^T$*.*

Expressing a symmetric matrix $\mathbf{A}$ in this form is called **diagonalizing the matrix**. We have shown that any symmetric matrix can be diagonalized so as to have its eigenvalues along the diagonal, and the matrix that achieves this is the matrix of eigenvectors.

Theorem 1.6.5. $\mathbf{A}$ *is negative (semi-)definite iff all of its eigenvalues are less than (or equal to)* 0*.*

Proof: $z^T\mathbf{A}z = z^T\mathbf{Q}^T\Lambda\mathbf{Q}z = v^T\Lambda v$, and the matrix $\mathbf{Q}$ is invertible. ∎

1.6.5. *The alternating signs determinant test for concavity.* Now we have enough matrix algebra background to prove what we set out prove, $\mathbf{A}$ is negative semi-definite (respectively negative definite) iff the sign of m'th principal sub-matrix is either 0 or -1^m (respectively, the sign of the m'th principal sub-matrix is -1^m).

We defined $g(y) = |\mathbf{A} - y\mathbf{I}|$ so that g is an n'th degree polynomial in λ, and used the fundamental theorem of algebra (and some calculation) to tell us that

$$g(y) = (\lambda_1 - y)(\lambda_2 - y)\cdots(\lambda_n - y)$$

where the λ_i are the eigenvalues of $\mathbf{A}$. Note that $g(0) = |\mathbf{A} - 0\mathbf{I}| = |\mathbf{A}| = \lambda_1 \cdot \lambda_2 \cdots \lambda_n$, that is,

Lemma 1.6.6. *The determinant of a matrix is the product of its eigenvalues.*

We didn't use symmetry for this result.

Recall that the **principal sub-matrices** of a symmetric $n \times n$ matrix $\mathbf{A} = (a_{ij})_{i,j=1,\dots,n}$ are the $m \times m$ matrices $(a_{ij})_{i,j=1,\dots,m}$, $m \leq n$. The following is pretty obvious, but it's useful anyway.

Problem 1.18. $\mathbf{A}$ *is negative definite iff for all* $m \leq n$ *and all non-zero* x *having only the first* m *components not equal to* 0, $x^T \mathbf{A} x < 0$.

Looking at $m = 1$, we must check if

$$(x_1, 0, 0, \dots, 0) \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{pmatrix} x_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = a_{11} x_1^2 < 0.$$

This is true iff the first principal sub-matrix of $\mathbf{A}$ has the same sign as $-1^m = -1^1 = -1$.

Looking at $m = 2$, we must check if

$$(x_1, x_2, 0, \dots, 0) \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} < 0.$$

This is true iff the matrix

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

is negative definite, which is true iff all of its eigenvalues are negative. There are two eigenvalues, the product of two negative numbers is positive, so the $m = 2$ case is handled by having the sign of the determinant of the 2×2 principal submatrix being -1^2.

Looking at $m = 3$, we must check if

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

is negative definite, which is true iff all of its eigenvalues are negative. There are three eigenvalues, the product of three negative numbers is negative, so the $m = 3$ case is handled by having the sign of the determinant of the 3×3 principal submatrix being -1^3.

Continue in this fashion, and you have a proof of Theorem 1.6.2. Your job is to fill in the details for the negative semi-definite, the positive definite, and the positive semi-definite cases as well.

Problem 1.19. *Prove Theorem* 1.6.2.

2. Some Basic Results in Metric Spaces

This section will cover continuity, compactness, the Theorem of the Maximum, and the Separating Hyperplane Theorem.

Readings: Kolmogorov and Fomin, intro to metric spaces, Ch. 2, sections 5-8.1, pp. 37-71.

Readings: Kolmogorov and Fomin, more on metric spaces, Ch. 3, sections 10-11.4, pp. 92-104.

Readings: Mas-Colell, Whinston, and Green on the Theorem of the Maximum, support functions and the supporting hyperplane theorem.

Readings: Sheldon Ross, Tables 1 and 2, p. 4.

2.1. **Metrics.** Let X be a set. This could be $\mathbb{R}^\ell$ or a subset thereof, $\mathbb{R}^{\mathbb{N}}$ or a subset (e.g. the bounded sequences, or ℓ_2) the set of probability distributions on $\mathbb{R}$ or $\mathbb{R}^\ell$ or a subset thereof, the set of convex preferences on $\mathbb{R}^\ell$ or a subset thereof, the set of possible demand functions or a subset (e.g. the set of differentiable demand functions), the set of technologies satisfying some natural set of restrictions, the set of budget sets or a subset thereof.

Definition 2.1.1. *A metric on X is any function $d : X \times X \to \mathbb{R}_+$ such that*

1. $d(x, y) = d(y, x)$,
2. $d(x, y) = 0$ *iff* $x = y$, *and*
3. $d(x, y) + d(y, z) \geq d(x, z)$.

In checking that something is a metric, the hardest part is usually the last inequality, known as the triangle inequality. For $r > 0$, $B(x, r) := \{y \in X : d(x, y) < \epsilon\}$. Draw some pictures. Kolmogorov and Fomin give many metrics on many spaces, read about them, figure out what the $B(x, r)$ "look like."

We will mostly use convergence of sequences. A sequence in X is a point in $X^{\mathbb{N}}$, denoted $(x_n)_{n\in\mathbb{N}}$ or x_n.

Dfn $x_n \to x$. Pictures about tail behavior should come to mind. Dfn closed set, open set, there are sets that are neither open nor closed.

2.2. **Probability distributions as cdf's.** We introduce a metric on cdf's as a metric on $\Delta(\mathbb{R})$, the set of probability distributions on $\mathbb{R}$. This is a precursor to prob/stats material involing the CLT.

Dfn cdf. These give probabilities on the field $\mathcal{F}_0$ generated by $\{(-\infty, a] : a \in \mathbb{R}\}$ by addition. We assume continuity from above at the empty set, which in our case is the same as right continuity (e.g. of problems is $F(x) = \frac{1}{2}1_{\{0\}}(x) + 1_{(0,\infty)}(x)$, $A_n = (0, \frac{1}{n}]$, $P_F(A_n) = F(\frac{1}{n}) - F(0) \equiv \frac{1}{2}$ but $\cap_n A_n = \emptyset$).

A metric on cdf's is

$$\rho(F,G) = \inf\{\epsilon > 0 \quad : \quad \forall x \in \mathbb{R},\ G(x) \leq F(x+\epsilon) + \epsilon \\ \text{and } F(x) \leq G(x+\epsilon) + \epsilon\}.$$

Levy ribbons and the triangle inequality. Weak* convergence aka convergence in distribution turns out to be equivalent to ρ-convergence. Let F_∞ be the cdf of δ_0, look at $B(F,\epsilon)$, let Φ be the standard normal cdf, look at $B(\Phi,\epsilon)$.

Example: X_t iid ± 1 $\frac{1}{2}$ each, F_T the cdf of $S_T := T^{-1}\sum_{t\leq T} X_t$, T **an <u>even</u> number**. By Tchebyshev we can show that for all $\epsilon > 0$, $P(|S_T - 0| > \epsilon) \to 0$, let F_∞ be the cdf of δ_0, and note that $F_T(0) \not\to F_\infty(0)$. Rather, $F_T(0) \equiv \frac{1}{2}$.

Dfn continuity point of F using sequences.

Dfn $F_n \to_{weak} F$ iff $F_n(x) \to F(x)$ for all continuity points of F. Compare CLT.

Theorem 2.2.1. *$F_n \to_{weak} F$ iff $\rho(F_n, F) \to 0$.*

2.3. **Continuity.** Metrics, τ_X is the collection of open sets, closedness, closed subsets of complete metric spaces are complete. Dfn: $f : X \to Y$ is cts if $f^{-1}(\tau_Y) \subset \tau_X$. Lemma: cts iff f^{-1} of the closed sets is a subset of the closed sets iff ϵ-δ iff sequence definition of ctuity.

Equality of topologies with different metrics, completeness does not survive change of metrics, $\rho(x,y) = |F(x) - F(y)|$, $F(r) = e^r/(1+e^r)$.

2.4. **Compactness and the existence of optima.**

1. **Heine-Borel** If $[a,b] \subset \cup_{\alpha\in A}(r_\alpha, s_\alpha)$, then $\exists$ finite $A_F \subset A$, $[a,b] \subset \cup_{\alpha\in A_F}$. Pf.
2. **Finite intersection property (fip)** A collection $\{F_\alpha : \alpha \in A\}$ of closed subsets of $[a,b]$ has the finite intersection property (fip) if $\cap_{\alpha\in A_F} F_\alpha \neq \emptyset$ for all finite $A_F \subset A$. Any collection of closed sets in $[a,b]$ with the fip satisfies $\cap_{\alpha\in A} F_\alpha \neq \emptyset$. By DeMorgan, this is equivalent to Heine-Borel.
3. If $f : [a,b] \to \mathbb{R}$ is cts, then $\exists x^* \in [a,b]$, $f(x^*) \geq f([a,b])$. Pf: By finite subcover of $f([a,b]) \subset \cup_{n\in\mathbb{Z}}(n, n+2)$, $f([a,b])$ is bounded, hence has a supremum, call it $\overline{f}$. The collection $\{f^{-1}([r - \frac{1}{n}, r]) : n \in \mathbb{N}\}$ has the fip, therefore $x^* \in \cap_n f^{-1}([r - \frac{1}{n}, r] \neq \emptyset$.
4. Turn Heine-Borel/fip into a dfn for metric spaces. Closed subsets of compact metric spaces are compact. Cts functions on compact metric spaces achieve their maximum.

2.5. **The Theorem of the Maximum.** For E a subset of X, define $E^\epsilon = \cup_{x\in E} B(x,\epsilon)$, this is the ϵ-ball around the set E. For compact $A, B \subset X$, define $m(A,B) = \inf\{\epsilon > 0 : A \subset B^\epsilon\}$. The Hausdorff distance between compact sets is

$$d(A,B) = \max\{m(A,B), m(B,A)\}.$$

Draw some pictures.

Definition 2.5.1. *A compact-valued, non-empty-valued correspondence* $\Gamma : X \longrightarrow Y$ *is*

1. **upper hemicontinuous (uhc) at** x *if for all* $\epsilon > 0$ $\exists \delta > 0$ $[x' \in B(x, \delta)] \Rightarrow [m(\Gamma(x'), \Gamma(x)) < \epsilon]$,
2. **uhc** *if it is uhc at all* x,
3. **lower hemicontinuous (lhc) at** x *if for all* $\epsilon > 0$ $\exists \delta > 0$ $[x' \in B(x, \delta)] \Rightarrow [m(\Gamma(x), \Gamma(x')) < \epsilon]$,
4. **lhc** *if it is lhc at all* x,
5. **continuous (cts) at** x *if it is both uhc and lhc at* x, *and*
6. **cts** *if it is cts at all* x.

Explosions of a correspondence can be uhc but not lhc, implosions of a correspondence can be lhc but not uhc.

A single valued Γ is uhc iff it is a continuous function.

Let $\mathcal{K}(Y)$ denote the compact subsets of Y. Γ is cts iff it is cts when viewed as a function from X to $\mathcal{K}(Y)$.

Theorem 2.5.2 (Theorem of the Maximum)**.** *If* $u : X \times Y \to \mathbb{R}$ *is cts and* $\Gamma : X \longrightarrow Y$ *is cts, compact and non-empty valued, then* $v(x) := \max\{u(x, y) : y \in \Gamma(x)\}$ *is a cts function and* $x \mapsto \{y : u(x, y) \geq u(x, \Gamma(x))\}$ *is an uhc correspondence.*

Applications: consumer choice theory, producer theory, general equilibrium, game theory. We will see this theorem in dynamic programming too.

2.6. **The Separating Hyperplane Theorem.** Hyperplanes, separation, the theorem, the proof. Applications: the 2'nd Welfare Theorem (existence of prices), existence of Lagrange multipliers [pass through saddle points and the simplest form of the Kuhn-Tucker theorem]. We did the basic duality theorem with applications to the recovery of preferences and technology from demand and supply behavior.

2.7. **Problems.**

Problem 2.1. $1_A(x)$ *is the indicator function of a set* A, *taking the value* 1 *when* $x \in A$ *and taking the value* 0 *otherwise. Show that* (X, ρ) *is a metric space when* X *is non-empty and* $\rho(x, y) = 1_{\{x \neq y\}}(x, y)$.

Problem 2.2. *The closure of a subset* E *of a metric space* (X, d) *is denoted* $\overline{E}$, *and is defined as the smallest closed set containing* E. *Show that the following are equivalent definitions of* $\overline{E}$*:*

1. $\overline{E} = \bigcap\{F : E \subset F,\ F \text{ closed}\}$.
2. $\overline{E} = \bigcap\{E^\epsilon : \epsilon > 0\}$.
3. $\overline{E} = \{x \in X : \forall \epsilon > 0, \exists e \in E, d(x, e) < \epsilon\}$.

4. $\overline{E} = \{x \in X : \exists e^n \text{ in } E \; e^n \to x\}$.

Since the intersection of any collection of closed sets is another closed set, the first really does capture the notion of the "smallest closed set containing E."

Problem 2.3. *Show that if $K \subset \mathbb{R}^\ell$ is convex, then $\overline{K} = \bigcap\{H_p^{\leq}(r) : K \subset H_p^{\leq}(r), p \in \mathbb{R}^\ell, r \in \mathbb{R}\}$.*

Problem 2.4. *In showing that $f : X \to \mathbb{R}$ achieves its maximum when f is continuous and (X, d) is compact, we used the fip. Reformulate this part of the proof using the open cover argument and the definition of a supremum.*

Problem 2.5. *Define a preference relation $\succsim$ on $\mathbb{R}^\ell_+$ to be* **continuous** *if for all $y \in \mathbb{R}^\ell_+$, the sets $\{x \in \mathbb{R}^\ell_+ : x \succsim y\}$ and $\{x \in \mathbb{R}^\ell_+ : y \succsim x\}$ are closed. Prove that for every non-empty, compact $K \subset \mathbb{R}^\ell_+$, $\exists x^* \in K$ such that $x^* \succsim K$.*

Problem 2.6. *(X, d) is compact iff for all sequences in X, $\exists x^\circ \in X$, $\exists$ a subsequence x_{n_k} such that $x_{n_k} \to x^\circ$.*

Problem 2.7. *Show that the closure of the ball around 0 with radius ϵ is compact in $\mathbb{R}^\ell$. Show that the closure of the ball around 0 with radius ϵ is <u>NOT</u> compact in $C_{[a,b]}$.*

Problem 2.8. *K-M,* §5.2, #1, #8 (p. 45).

Problem 2.9. *K-M,* §5.2, #3, 4, 5 (p. 54).

Problem 2.10. *K-M,* §5.2, #9, 10 (p. 54-5).

Problem 2.11. *For each n, let X_{nk}, $k = 1, \ldots, n$, be independently distributed Bernoulli(λ/n) and define $Y_n = \sum_{k \leq n} X_{nk}$. Let F_n be the cdf of Y_n. Show that $F_n \to_{weak} F_\lambda$ where F_λ is the cdf of a Poisson(λ) distribution.*

3. Dynamic Programming, Deterministic and Stochastic

Readings: Ross, Chapters 1, 2.1-2, 2.4, 4.1-3, and 6.1-5.

Optimization in dynamic contexts is more difficult, and substantive results that are true in generality are hard to come by. By contrast, in neoclassical demand theory, the general result is that demand functions have a negative semi-definite Slutsky matrix. Don't expect anything so definite here without piles of extra assumptions.

3.1. **Compactness and continuity in spaces of sequences.** In consumer demand theory, $\max\ u(x)$ s.t. $x \geq 0$, $px \leq m$ has a solution if u is continuous because the constraint set is compact. There is a parallel result for dynamic programming.

Each (Y_t, d_t) compact, $t = 0, 1, \ldots$, $Y := \times_t Y_t$, the metric on Y is

$$d(x, y) := \sum_t 2^{-t} \min\{1, d_n(x_t, y_t)\}.$$

Theorem 3.1.1. *(Y, d) is compact.*

Proof: Diagonalization. ∎

If $u : Y \to \mathbb{R}$ is d-cts, then it is asymptotically tail-insensitive, indeed, to within any $\epsilon > 0$, a cts u depends ctsly on only finitely many coordinates. To say this more precisely, for $x, y \in Y$ and $t \in \mathbb{N}$ let $x/_t y \in Y$ be the point (i.e. sequence) which is x up to and including time t and is y thereafter, i.e.

$$x/_t y = (x_0, x_1, x_2, \ldots, x_{t-1}, x_t, y_{t+1}, y_{t+2}, \ldots).$$

Note that many of the ways we valued sequences of rewards are **NOT** continuous, e.g. $V_{\liminf}(x)$ is discontinuous — let $\tilde{r}$ denote the sequence $(r, r, r, r, \ldots)$, for any x and any r, $x/_t \tilde{r} \to_t x$, and for all t, $V_{\liminf}(x/_t \tilde{r}) = r$. More specifically, Let $x = (0, 1, 0, 1, 0, 1, 0, 1, 0, 1, \ldots)$, $V_{\liminf}(x) = 0$, $(x/_t \tilde{r}) \to_t x$, and $V_{\liminf}(x/_t \tilde{r}) \not\to 0$ unless $r = 0$.

Theorem 3.1.2. *If $u : Y \to \mathbb{R}$ is continuous, then for all $\epsilon > 0$, $\exists T\ \forall t \geq T$,*

$$\max\{|u(x/_t y) - u(x/_t y')| : x, y, y' \in Y\} < \epsilon.$$

Proof: Uniform continuity and shape of the d-balls. ∎

3.2. **Deterministic Dynamic Programming.** Deterministic dynamic programming problems that economists use are almost always problems that maximize continuous functions over compact subsets of compact Y's of the form given. They almost always have the following structure: (i) one starts with x_0, picks x_1, then restarts with x_1 and picks x_2, and on and on; (ii) the possible x_t are constrained by x_{t-1}; (iii) in each period t, the reward depends on the "state," x_{t-1} and the action, x_t; (iv) rewards are added across periods and discounted; (v) this period's action becomes next period's state.

Some notation, $(x_0, x_1, x_2, \dots) = (x_0, x_{1+})$.

1. $Y_n \subset X$ for some metric space X,
2. $\exists \varphi : X \times X \to \mathbb{R}$ cts such that

$$u(x_0, x_{1+}) = \varphi(x_0, x_1) + \sum_{t \geq 1} \beta^t \varphi(x_t, x_{t+1})$$

for some $\beta \in (0, 1)$, and

3. $x_t \in \Gamma(x_{t-1})$ for some continuous, compact-valued correspondnce Γ from X to X.

Some assumptions are needed on the relation of φ and Γ to guarantee that $u(\cdot)$ is cts, the simplest is that φ be cts and bounded, another, less "primitive" assumption is summability for the largest possible $\varphi(x_t, x_{t+1})$'s, the leading example of which has the maximal growth rate eventually a fraction of the discount rate.

If there's time, discuss the finite horizon value function.

The basic fish/tree growth model, intuitions. The value function for the infinite horizon case. $\Gamma_\infty : X \longrightarrow X^{\mathbb{N}}$, $\Gamma_\infty(x_0) := \{(x_t)_{t \geq 1} : x_1 \in \Gamma(x_0), x_{t+1} \in \Gamma(x_t), t = 1, 2, \dots\}$, $U(x_0, (x_t)) := \varphi(x, x_1) + \sum_{t \geq 1} \beta^t \varphi(x_t, x_{t+1})$. As Γ_∞ is a cts compact-valued correspondence and $U : X \times X^{\mathbb{N}} \to \mathbb{R}$ is a cts function, $V(x_0) := \max\{U(x_0, (x_t)) : (x_t) \in \Gamma_\infty(x_0)\}$ is cts (by the Theorem of the Maximum), and the solution set is uhc.

Theorem 3.2.1. *For all $x \in X$, if $(x_t^*) \in \Gamma_\infty(x)$ satisfies $V(x) = U(x, (x_t^*))$, then $V(x) = \varphi(x, x_1^*) + \beta V(x_1^*)$ and for all t, $V(x_t^*) = \varphi(x_t^*, x_{t+1}^*) + \beta V(x_{t+1}^*)$. Further, $V(x_0) = U(x_0, (x_t'))$ iff for each $t \geq 1$, x_t' solves $\max\{\varphi(x_{t-1}, y) + \beta V(y) : y \in \Gamma(x_{t-1})\}$.*

A crucial implication is that once one has found V, the optimal policy can be found by

$$P^*(x) = \arg\max\{\varphi(x, y) + \beta V(y) : y \in \Gamma(x)\}.$$

While there may be many best ways to do something, that is, $P^*(x)$ may contain more than one point, the value of doing any one of the best things is unique. The remaining problem is how to find V. There is an approximation method based on the Contraction Mapping Theorem.

Let $C_b(X)$ denote the cts bdd functions on X. Define $d(f, g) = \sup_{x \in X} |f(x) - g(x)|$.

Lemma 3.2.2. *$(C_b(X), d)$ is a complete metric space.*

Proof: An $\epsilon/3$ argument. ∎

For any W in $C_b(X)$, define $\Psi(W)$ by

$$\Psi(W)(x) = \max\{\varphi(x, y) + \beta W(y) : y \in \Gamma(x)\}.$$

Interpret, e.g. $W \equiv 0$, interative applications of Ψ. Ψ is a contraction mapping from C_b to C_b. The content of the next theorem is that Ψ must have a unique fixed point. We care about this result since the unique fixed point turns out to be the value function.

Definition 3.2.3. *Let (Y,d) be a metric space. A function $f:Y\to Y$ is a* **contraction mapping** *if $\exists\beta\in(0,1)$, $\forall x,y\in Y$, $d(f(x),f(y))\leq\beta d(x,y)$.*

Any contraction mapping must be cts,

$$[d(y^n,y)\to 0]\Rightarrow[d(f(y^n),f(y))\leq\beta\cdot d(y^n,y)\to 0].$$

Lemma 3.2.4. *$\Psi:C_b(X)\to C_b(X)$ is a contraction mapping.*

Proof: $\Psi(W)\in C_b$ by the Theorem of the Maximum. The rest is the usual argument, it's in the text, and I'll give it in lecture. ∎

Since $C_b(X)$ is a complete metric , we can apply

Theorem 3.2.5 (Contraction Mapping)**.** *If $f:Y\to Y$ is a contraction mapping and Y is a complete metric space, then there exists a unique y^* such that $f(y^*)=y^*$. Further, for any $y_0\in Y$, the inductively defined sequence $y^n=f(y_{n-1})$ converges to y^*.*

Proof: Step 1 — if such a y^* exists, it is unique. To see why, suppose that $f(y^*)=y^*$ and $f(y')=y'$ so that $d(y^*,y')=d(f(y^*),f(y'))$. By the definition of a contraction mapping, $d(y^*,y')\leq\beta d(f(y^*),f(y'))$ for some $\beta<1$. Combining, $d(y^*,y')\leq\beta d(y^*,y')$, and this is only possible if $d(y^*,y')=0$.

Step 2 — existence. Pick an arbitrary $y\in Y$. Inductively define $f^0(y)=y$ and $f^n(y)=f(f^{n-1}(y))$. Applying the definition of a contraction mapping n times, we have

$$d(f^{n+m}(y),f^n(y))\leq\beta^n d(f^m(y),y).$$

Using the triangle inequality m times, we have

$$\beta^n d(f^m(y),y)\leq\beta^n[d(f^m(y),f^{m-1}(y))+\cdots+d(f(y),y)].$$

By the definition of a contraction mapping,

$$\beta^n[d(f^m(y),f^{m-1}(y))+\cdots+d(f(y),y)]\leq\beta^n d(f(y),y)[1+\beta+\cdots+\beta^{m-1}].$$

This last term, $\beta^n d(f(y),y)[1+\beta+\cdots+\beta^{m-1}]$, goes to 0 as $n\uparrow\infty$. Since Y is complete, there exits a y^* such that $y^*=\lim_n f^n(y)=\lim_n f^{n+1}(y)$. Because the function f is continuous, $f(y^*)=f(\lim_n f^n(y))=\lim_n f^{n+1}(y)=y^*$. ∎

Discuss starting at $W\equiv 0$ and applying Ψ.

3.3. **Stochastic Dynamic Programming.** Discrete Markov chains and Markovian dynamic programming (Ross, parts of Chapters 1, 2, 4, and 6).

Chapter 1. Random variables: as distributions, as functions on $([0,1],\mathcal{B},\lambda)$, as characteristic functions. State (but do not yet prove) uniqueness and convergence results for characteristic functions.

Conditional probabilities and expectations: discrete cases, $E(Y|X=x)$ is a function of x, $Y=1_A$ and $E(Y|X=x)=P(A|X=x)$. Stochastic process: we'll have discrete time processes.

Chapter 2. Poisson processes (star-finite dfns), interarrival and waiting time distributions (negative exponential and gamma), nonhomogenous and compound processes.

Chapter 4. Markov chains, Examples 2 and 3 of embedded Markov chains, classification of states, limit theorems, special emphasis on finite state spaces and the contraction mapping proof of Froebenius's theorem for positive matrixes.

Chapter 6. Markov decision processes with discrete state spaces and finite actions, a (triumphant) return to the contraction mapping. Application of these ideas to sequential testing.

3.4. **Problems.**

Problem 3.1. *Which, if any, of the criteria for valuing sequences given above,* V_1, V_2, $V^{average}_{\liminf}$, V_3, V_{Polya}, *are continuous? Prove your answers.*

Problem 3.2. *A complex number* z *is a vector in* $\mathbb{R}^2$. *For complex numbers* (u, v) *and* (x, y), *complex addition is defined by* $(x, y)+(u, v) = (x+u, y+v)$ *and complex multiplication by* $(x, y)(u, v) = (xu - yv, xv + yu)$. *The first component of* $z = (x, y)$ *is called its real part, the second component the imaginary part so that* $z = (x, 0)+(0, y)$ *expresses* z *as the sum of its real and imaginary parts. The complex conjugate of* $z = (x, y)$ *is defined as* $\overline{z} = (x, -y)$. *The absolute value of* $z = (x, y)$ *is defined by* $|z| = \sqrt{x^2 + y^n}$. *The complex number* i *is defined as* $(0, 1)$. *Typical notation is* $z = x + iy$ *where* $x, y \in \mathbb{R}$. *The more complete, clumsy notation is* $z = x(1, 0) + y(0, 1)$. *For* $z \neq (0, 0)$, *define* $1/z := \overline{z}/|z|$.

1. *Restricted to the real axis, complex addition and multiplication are the usual addition and multiplication.*
2. *Show that* $i^2 = -1$.
3. *Interpret* $z\overline{z}$.
4. *Show that* $1/z = (x/(x^2 + y^2), -y/(x^2 + y^2)$. *Interpret geometrically.*
5. *Any* z *can be identified with a pair* (r, θ) *of polar coordinates where* $r = |z|$ *and* θ *is the angle between* z *and the line segment starting at* $(0, 0)$ *and extending through* $(1, 0)$. *There is some ambiguity in this because* $\theta + 2n\pi$ *is also the angle for any* $n \in \mathbb{N}$. *We take the value in* $[0, 2\pi)$.
   (a) *Give the formula for coverting* $z = (x, y)$ *into* (r, θ).
   (b) *Give the polar coordinate formula for complex multiplication.*

Problem 3.3. *This problems asks you to go through the two major parts of the central limit theorem in their simplest forms.*

1. *For each* n, *let* X_{nk}, $k = 1, \ldots, n$, *be independently distributed Bernoulli*(λ/n). *Define* $Y_n = \sum_{k\leq n} X_{nk}$. *Calculate the characterstic function,* $\varphi_n(u)$ *of* Y_n, *and show that for all* u, $\varphi_n(u) \to \varphi_\lambda(u)$ *where* $\varphi_\lambda(\cdot)$ *is the characteristic function of a Poisson*(λ).

2. *For each n, let X_{nk}, $k = 1, \dots, n$, be independently distributed with $P(X_{nk} = -1) = P(X_{nk} = +1) = \frac{1}{2}$. Define $Y_n = \frac{1}{\sqrt{n}} \sum_{k \leq n} X_{nk}$. Calculate the characterstic function, $\varphi_n(u)$ of Y_n, and show that for all u, $\varphi_n(u) \to \varphi_{Gaussian}(u)$ where $\varphi_{Gaussian}(\cdot)$ is the characteristic function of the standard Gaussian distribution.*

Problem 3.4. *Ross,* Chapter 1, #1, 2, 4, 7, 8, 11.

Problem 3.5. *Ross,* Chapter 2, #1, 3, 7.

Problem 3.6. *Ross,* Chapter 4, #1, 2, 3, 5, 8, 12, 14.

Problem 3.7. *Ross,* Chapter 6, #1, 2, 3, 4, 5, 6.

4. An Overview of the Statistics Part of this Course

Concepts we will see ofter are marked with a "†."

4.1. **Basics.** The basic statistical model has data† $\mathbf{X} = (X_1, \ldots, X_n)$ independent†, and identically distributed with distribution P_θ, $\theta \in \Theta$. Given θ, the likelihood† of $\mathbf{X}$ is $L(\mathbf{X}|\theta) = \Pi_i P_\theta(X_i)$. Maximum likelihood estimators† (MLE's) solve the problem $\max_{\theta\in\Theta} L(\mathbf{X}|\theta)$ for $\widehat{\theta}(\mathbf{X})$. When the P_θ's all have densities, we use those. Generally, taking the logarithm of L makes the calculations easier.

There are many ways to arrive at sensible estimators. MLE's are not the only class of estimators that we will look at, but they are a very good starting point.

Any function of the data is a **statistic**†. Writing $\widehat{\theta}(\mathbf{X})$ makes it clear that estimators are statistics. We often solve for them as a function of the possible values that $\mathbf{X}$ may take on, that is, we solve for $\widehat{\theta}(\mathbf{x}) := (\widehat{\theta}(\mathbf{X})|\mathbf{X} = \mathbf{x})$, $\mathbf{x} = (x_1, \ldots, x_n)$.

As a first, and very informative example, suppose that the data, $X_1, \ldots, X_n$, are independent† random variables, distributed Bernoulli(p), $p \in [0, 1]$. We want $\widehat{p} = \widehat{p}(X_1, \ldots, X_n)$, an estimator† of p. From intro stats, $Prob((X_1, \ldots, X_n) = (x_1, \ldots, x_n)) = \Pi_i p^{X_i}(1-p)^{1-X_i}$, from this find the maximum likelihood estimator† (MLE), $\widehat{p}_n = \frac{1}{n}\sum_{i\le n} X_i$.

This is a linear statistic, and it satisfies, $\forall p \in [0, 1]$, conditional on p being the true value, we have the expectation of $\widehat{p}_n$ is p, written as $E(\widehat{p}_n|p) = p$, that is, it is unbiased†. Indeed, amongst the set of unbiased linear statistics, it minimizes variance. Further, by the Law of Large Numbers† (LLN), $Prob(\widehat{p}_n \to p) = 1$, that is, the estimator is consistent†.

Having an estimator is fine and lovely, one of my favorites is $\frac{1}{7}$. What we would like to know is $Prob(|\widehat{p}_n - p| > r)$ for different values of r and different values of p. More generally, we would like to know different aspects of the distribution of $\widehat{p}$ for different p. Graph the distributions for $\widehat{p} \equiv \frac{1}{7}$ and for the MLE.

In general, we would like to know about the likelihood that our estimator is very far off, and how that depends on the true value. We can do explicit calculations, or we can use the Central Limit Theorem† (CLT) for the mean, as you should remember. If $X_1, \ldots, X_n$ are iid with mean μ and variance σ^2, then $\frac{1}{\sqrt{n}}\sum_i (X_i - \mu)/\sigma$ is, to a good degree of approximation, distributed $N(0, 1)$.

4.2. **Other properties of estimators.** Suppose now that p is the true value, that we do not know it, and want an estimator $\widehat{p}'$ with low **mean squared error**†, (MSE) that is, such that $E((\widehat{p}' - p)^2|p)$ is small. It is a true fact that shrinking $\widehat{p}_n = \frac{1}{n}\sum_{i\le n} X_i$ lowers MSE. We'll do the algebra for non-negative statistics, passing through MSE = Var + Bias. Generally, the amount of shrinkage to be done to minimize MSE depends on the true value of the parameter we're interested. When we put our estimate of the parameter into the formula for the optimal shrinkage and shrink, we may actually **increase** the MSE.

The Best MSE estimators are biased, the better ones are called "shrunken" estimators. Suppose that $\hat{\theta} \in \mathbb{R}_{++}$ is an unbiased estimator of a location parameter $\theta \in \mathbb{R}_{++}$ based on an iid. sample X_i, $i = 1, \ldots, n$. The question to be asked is what multiple of $\hat{\theta}$ minimizes mean squared error? To answer the question, we take a detour through the following calculation:

$$E(\hat{\theta} - \theta)^2 = E((\hat{\theta} - E\hat{\theta}) + (E\hat{\theta} - \theta))^2 \tag{3}$$

$$= E(\hat{\theta} - E\hat{\theta})^2 + E(E\hat{\theta} - \theta)^2 + 2E(\hat{\theta} - E\hat{\theta})(E\hat{\theta} - \theta) \tag{4}$$

$$= \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}^2(\hat{\theta}), \tag{5}$$

where the last equality happens because $E\hat{\theta}$ is unbiased.

We now apply this to the class of estimators $a\hat{\theta}$ where $\hat{\theta}$ is unbiased. Define

$$f(a) = E(a\hat{\theta} - \theta)^2 = a^2\mathrm{Var}(\hat{\theta}) + \theta^2(a-1)^2.$$

Let $v = \mathrm{Var}(\hat{\theta})$, so that $f(a) = a^2 v + \theta^2(a-1)^2$. Because f is a quadratic in a with positive coefficients on a^2, the first order conditions are sufficient for a maximum. Taking derivatives, $\frac{1}{2}f'(a) = av + \theta^2(a-1)$ so that

$$a^* = \frac{\theta^2}{v + \theta^2} = \frac{1}{1 + v^\circ} < 1,$$

where $v^\circ := \frac{v}{\theta^2}$ (which is known as the standardized variation of $\hat{\theta}$). Thus, the optimal MSE (Mean Squared Error) estimator which is a linear function of $\hat{\theta}$ is given by $a^*\hat{\theta}$. Because $a^* < 1$, these are sometimes called shrunken estimators.

Note that as v° becomes large, a^* becomes small. As $n \uparrow \infty$, $v^\circ \downarrow 0$ for the estimators that we will study, so that as more data arrives, a^* approaches 1. It can be shown that for the negative exponential distribution, $a^* = \frac{n}{n+1}$. In general, one must estimate v° to figure out what a^* should be, so the resulting estimate is $\hat{a}^*\hat{\theta}$. One could avoid thinking about this problem simply by insisting on unbiased estimators, this has the effect of forcing $a^* = 1$, but there doesn't seem to be much justification for that. However, in some cases, $\hat{a}^*\hat{\theta}$ turns out to have even higher MSE than $\hat{\theta}$ did originally. (Onwards, bravely, through the fog.)

4.3. **Bayesians.** Another way to look at the whole problem gives very direct answers to questions about the value of $Prob(|\widehat{p}_n - p| > r)$ and, more generally, the distribution of $\widehat{p}_n$. This approach is called Bayesian statistics†, and, at its best, it sensibly uses the prior knowledge we have about the problem at hand.

Suppose that we know that the true value of θ (changing the notation for p here) is in the interval $[\frac{1}{2}, 1]$, and that intervals of equal size in $[\frac{1}{2}, 1]$ are equally likely, that is, our prior distribution† is $U[\frac{1}{2}, 1]$. The posterior density as a function of the data is

$$P(\theta|\mathbf{x}) = k_x\theta \cdot \Pi_i \theta^{X_i}(1-\theta)^{1-X_i}, \quad \theta \in [\tfrac{1}{2}, 1],$$

where k_x is some constant chosen to make $\int_{[\frac{1}{2},1]} kP(\theta|\mathbf{x})\,d\theta = 1$. Take logarithms and maximize over $[\frac{1}{2}, 1]$ to find the Bayesian MLE estimator, watch out for corner solutions.

One of the fascinating aspects of the study of statistics is the interplay between the ideas implicit in the MLE approach and this Bayesian approach. The basic issues in epistemology appear, how do we know the information in the prior? And how sure of it are we? One way to answer these questions appear when there is a lot of data. In this case, there is a tight relation between Bayesian estimators and MLE estimators.

Suppose that $\theta \in \Theta$, and a Bayesian has a prior distribution with density $p(\theta)$, and we observe $X_1, \ldots, X_n$ with density $f(x|\theta)$. Then the posterior distribution has density

$$P(\theta|X_1, \ldots, X_n) = kp(\theta)L(X_1, \ldots, X_n|\theta)$$

for some constant k. A Bayesian might well solve the problem $\max_\theta P(\theta|X_1, \ldots, X_n)$. Taking logarithms, this gives

$$\max_\theta[\log p(\theta) + \textstyle\sum_i \log f(X_i|\theta)] = \max_\theta \sum_i [\log f(X_i|\theta) + \frac{1}{n}\log p(\theta)].$$

You should be able to convince yourself that the solution to this problem approaches the MLE as $n \uparrow \infty$. We interpret this as saying that the prior distribution becomes irrelevant, it is eventually swamped by the data. For moderate n, the approximation may not be that good.

4.4. **Classical statistics.**

> Hypothesis — A supposition or conjecture put forth to account for known facts; esp. in the sciences, a provisional supposition from which to draw conclusions that shall be in accordance with known facts, and which serves as a starting-point for further investigation by which it may be proved or disproved and the true theory arrived at. (OED)

Hypothesis testing involves formulating a supposition, or conjecture, or guess, called the "null hypothesis," written H_0, and the "alternative hypothesis," H_1 or H_A, then observing data that has different distributions under H_0 and H_1, and then picking between the two hypotheses on the basis of the data. Under study are the possible processes of picking on the basis of the data. This is formulated as a decision rule, aka a rejection rule, that is, for some set of possible data points we reject H_0, for others we accept it.

There are two types of errors that a rejection rule can make, unimaginatively called **Type I** and **Type II** errors:

1. you can reject a null hypothesis even though it is true, this kind of false rejection of a true null hypothesis is called a **Type I** error, the probability of a Type I error is denoted α.

2. you can accept the null hypothesis even though it is false, this kind of false acceptance of a null hypothesis is called a **Type II** error, the probability of a Type II error is denoted β.

If you adopt a rejection rule that makes α small, you are very rarely rejecting the null hypothesis. This means that you are running a pretty high risk of making a Type II error, that is, β is fairly large. This works the other way too, getting β small requires accepting a large α. The Neyman-Pearson Lemma concerns a class of situations in which we can find the best possible decision rule for given α.

The essential ingredients are

1. The basic statistical model, $\mathbf{X} \sim f(\mathbf{x}|\theta)$, $\theta \in \Theta$,
2. a null hypothesis, $H_0 : \theta \in \Theta_0$, $\Theta_0 \subset \Theta$,
3. the alternative, $H_1 : \theta \notin \Theta_0$,
4. a decision rule, reject H_0 if $\mathbf{x} \in \mathbb{X}_r$ and accept H_0 if $\mathbf{x} \notin \mathbb{X}_r$.

We can then examine the probabilistic properties of the decision rule using the **power function**, $\beta(\theta) = P(\mathbb{X}_r|\theta)$. The perfect power function is $\beta(\theta) = 1_{\Theta_0^c}(\theta)$, that is, reject if and only if the null hypothesis is false. (Sing a bar of "To dream the impossible dream.") However, the idea behind the basic statistical model is that we do not observe θ directly, rather we observe the data $\mathbf{X}$, and the data contains probabilistic information about θ. In statistics, we don't expect to see perfect power functions, they correspond to having positive proof or disproof of a null hypothesis.

Continuing in our simplest of examples, we suppose $H_0 : p = p_0$, $H_1 : p = p_1$, $p_0 \neq p_1$. Notice how much structure we've already put on the problem of picking a decision rule. We'll suppose that $p_0 < p_1$, the opposite case just reverses inequalities. There is a pretty strong intuition that the best decision rule is to accept H_0 if $\widehat{p}_n < p^*$, and to reject otherwise, for some $p^* \in (p_0, p_1)$. Work through why this is true, thinking of the analogy with filling bookcases with the lightest possible set of books.

4.5. **An Information Inequality.** We saw the Cauchy-Schwarz inequality for vectors, $xy = \|x\|\|y\| \cos\theta$, equivalently, $\sum_i x_i y_i = \sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2} \cos\theta$, so that $(\sum_i x_i y_i)^2 \leq \sum_i x_i^2 \sum_i y_i^2$. When $\Omega = \{1, \ldots, n\}$ with $P(\omega) \equiv 1/n$, we get an inequality about expectations that is also an inequality about Variances and Covariances,

$$\mathrm{Cov}(X, Y)^2 \leq \mathrm{Var}(X)\mathrm{Var}(Y)$$

with equality iff $X - E\,X$ is linear function of $Y - E\,Y$.

Go through definition $\mathrm{Cov}(X, Y) = E(X - E\,X)(Y - E\,Y) = E\,XY - E\,X E\,Y$. Defining the L^2 norm, and $\min_\lambda \|X - \lambda Y\|$ for a more general version.

Now, $\mathrm{Var}(Y) > 0$ iff $P(Y \neq EY) > 0$, so division yields

$$\mathrm{Var}(X) \geq \frac{\mathrm{Cov}(X,Y)^2}{\mathrm{Var}(Y)}$$

for all of the interesting Y's.

We are going to take $R = \widehat{p}$, $S = D_p \log f(\mathbf{X}|p)$ and look at

$$\mathrm{Var}_p(R) \geq \frac{\mathrm{Cov}_p(R,S)^2}{\mathrm{Var}_p(S)}.$$

Here $E_p X$ is the expectation of the rv X when the true value is p. We will show that when R is unbiased, $\mathrm{Cov}_p(R,S)^2 = 1$. We will also show that $E_p(D_p \log f(\mathbf{X}|p)) = 0$. This means that for **any** unbiased estimator, $\hat{p}$, of p,

$$\mathrm{Var}_p(\hat{p}) \geq \frac{1}{E_p(D_p \log f(\mathbf{X}|p))^2},$$

known as the **Cramèr-Rao lower bound**. It really is a bound on all unbiased estimators, the right hand side does not depend on which $\hat{p}$ you choose. If we have an unbiased estimator where this inequality is satisfied as an equality, then we have found the smallest possible variance amongst all unbiased estimators. Sometimes there is no estimator satisfying the bound.

The quantity $E_p(D_p \log f(\mathbf{X}|p))^2$ is called the **Fisher information** of the sample, and the inequality, in this form, is often called the **information inequality**. It should be intuitive that having $E_p(D_p \log f(\mathbf{X}|p))^2$ large means that the sample tells us a great about the value of the parameter, that we can estimate it more closely. This is especially true when we think about the MLE's.

4.6. **Mis-Specification.** The basic statistical model is $X_1, \ldots, X_n$ is i.i.d. P_θ for some $\theta \in \Theta$. Any subset of the parts of the model may be incorrect.

1. Independence is typically violated in time series contexts, e.g. $X_{t+1} = f(X_t) + \epsilon_t$ and $\mathrm{Cov}(\epsilon_t, \epsilon_{t+1}) \neq 0$. Think about panel data sets in labor.
2. Identical distribution may be violated, e.g. p is lower for the first half than for the second half of the treatment because of learning on the part of the administrators. Errors for predictive equations may be systematically smaller in absolute value for some identifiable part of the sample.
3. If $X_1, \ldots, X_n$ is iid μ and $\mu \notin \{P_\theta : \theta \in \Theta\} \subset \Delta$.
   (a) Random X_i's and X_i''s, $Y_i = \beta_0 + \beta_1 X_i + \beta_1' X_i' + \epsilon_i$, ϵ_i iid $N(0, \sigma^2)$, X_i, X_i' iid some distribution. We are interested in estimating the β's and σ^2. Suppose we use the model $Y_i = \beta_0 + \beta_1 X_i + +\epsilon_i$, ϵ_i iid $N(0, \sigma^2)$, X_i iid some distribution. It's not hard to write down the MLE from the mis-specified model, and if $\mathrm{Cov}(X_i, X_i') \neq 0$, the

estimator of β_1 is biased (famous cases, fertilizer and yield, education and wages). The good properties of MLE's need not survive mis-specification.

(b) The X_i's in the above model might not be random, they could have been designed (not generally true in econ), or might not be replicable (generally true in econ), in which case our analysis/estimators are conditional on the data rather than having more generalized good properties.

4.7. **Problems.**

Problem 4.1. *Suppose that $X_1, \ldots, X_n$ are iid, $E\,X_i = \mu$, $Var(X_i) = \sigma^2$. A linear estimator of μ is a function of the form $\widehat{\mu}_\alpha(X_1, \ldots, X_n) = \sum_i \alpha_i X_i$, $\alpha = (\alpha_1, \ldots, \alpha_n)$. An estimator, $\widehat{\mu}$, is unbiased if for all μ, $E(\widehat{\mu}|\mu) = \mu$.*

1. *Characterize the set of unbiased linear estimators.*
2. *Amongst the unbiased linear estimators, find the one with the lowest variance.*

Problem 4.2. *Suppose that $X_1, \ldots, X_n$ are iid Exponential(β), $\beta \in \mathbb{R}_{++}$.*

1. *Find $\widehat{\beta}_{MLE}(X_1, \ldots, X_n)$ and show that it is unbiased.*
2. *Show that the optimal shrinkage for $\widehat{\beta}_{MLE}$ is $\frac{n}{n+1}$, independent of β.*

Problem 4.3. *Suppose that $X_1, \ldots, X_n$ are iid Poisson(λ), $\lambda \in \mathbb{R}_{++}$.*

1. *Find $\widehat{\lambda}_{MLE}(X_1, \ldots, X_n)$ and show that it is unbiased.*
2. *Show that the optimal shrinkage for $\widehat{\lambda}_{MLE}$ depends on λ.*

Problem 4.4 (Neyman-Pearson)**.** *Suppose that $\mathbf{X} = (X_1, \ldots, X_n)$ has pdf (or pmf, in which case the integrals below are replaced by sums) $f(\mathbf{x}|\theta)$, $\theta \in \Theta = \{\theta_0, \theta_1\}$. We have seen that there is typically a tradeoff between α, the probability of a Type I error, and β, the probability of a Type II error. Let us suppose that we dislike both types of errors, and in particular, that we are trying to devise a test, characterized by its rejection region, $\mathbb{X}_r$, to minimize*

$$a \cdot \alpha(\mathbb{X}_r) + b \cdot \beta(\mathbb{X}_r)$$

where $a, b > 0$, $\alpha(\mathbb{X}_r) = P(\mathbf{X} \in \mathbb{X}_r|\theta_0)$, and $\beta(\mathbb{X}_r) = P(\mathbf{X} \notin \mathbb{X}_r|\theta_1)$. The idea is that the ratio of a to b specifies our tradeoff between the two Types of error, the higher is a relative to b, the lower we want α to be relative to β. This problem asks about tests of the form

$$\mathbb{X}_{a,b} = \{\mathbf{x} : af(\mathbf{x}|\theta_0) < bf(\mathbf{x}|\theta_1)\} = \left\{\mathbf{x} : \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} > \frac{a}{b}\right\}.$$

*This decision rule is based on the **likelihood ratio**, and likelihood ratio tests appear regularly in statistics.*

1. *Show that a test of the form $\mathbb{X}_{a,b}$ solves the minimization problem given above. [Hint: let $\phi(\mathbf{x}) = 1$ if $\mathbf{x} \in \mathbb{X}_r$ and $\phi(x) = 0$ otherwise. Note that $a \cdot \alpha(\mathbb{X}_r) + b \cdot \beta(\mathbb{X}_r) =$*

$a\int \phi(\mathbf{x})f(\mathbf{x}|\theta_0)\,d\mathbf{x}+b\int(1-\phi(\mathbf{x}))f(\mathbf{x}|\theta_1)\,d\mathbf{x}$, *and this is in turn equal to* $b+\int \phi(\mathbf{x})[af(\mathbf{x}|\theta_0)-bf(\mathbf{x}|\theta_1)]\,d\mathbf{x}$. *The idea is to minimize the last term in this expression by choice of* $\phi(\mathbf{x})$. *Which* $\mathbf{x}$*'s should have* $\phi(\mathbf{x})=1$*?*]

2. *As a function of* a *and* b, *find the* $\mathbb{X}_{a,b}$ *when* $(X_1,\ldots,X_n)$ *is iid Bernoulli*(θ), $\theta\in\Theta=\{\theta_0,\theta_1\}\subset(0,1)$.

Problem 4.5. *Suppose that* $X_1,\ldots,X_n$ *are iid Poisson*(λ), $\lambda\in\mathbb{R}_{++}$. *Show that* $\widehat{\lambda}_{MLE}(\mathbf{X})$ *achieves the Cramèr-Rao lower bound.*

5. Basic Probability, Transformations, and Expectations

5.1. **Basic Probability and Expectations.** Dfn $(\Omega, \mathcal{F}, P)$. Countable additivity as continuity from above at the empty set. On $\mathbb{R}$, cdf's and pdf's or pmf's, the fundamental theorem of calculus, the uniqueness of prob's from cdf's, look through the Table of Common distributions, p. 621-7, finding the expectation of X from F_X. Properties of σ-fields, the role of $[E_n$ i.o.] and $[E_n$ a.a.], DeMorgan's Rules and the complements of $[E_n$ i.o.] and $[E_n$ a.a.]. Conditional probabilities and independence, disjointness versus independence, Bayes Law and legal theory, the Monte Hall story.

5.2. **Transformations and Expectations.** Chain rule and $F_Y(y)$, $Y = g(X)$, g monotonic, g not monotonic, Leibniz's rule, applications from Chapter 2.

Differentiating under the integral sign. From Lebesgue's Dominated Convergence Theorem — if $(x, y) \mapsto h(x, y)$ is continuous at y_0 for each x and $|h(x, y)| \leq g(x)$ for some $g(x)$ satisfying $\int_{\mathbb{R}} |g(x)|\, dx < \infty$, then

$$\lim_{y \to y_0} \int_{\mathbb{R}} h(x, y)\, dx = \int_{\mathbb{R}} \lim_{y \to y_0} h(x, y)\, dx.$$

Corollary, if $(x, \theta) \mapsto f(x, \theta)$ is differentiable at θ_0 for every x, that is,

$$\lim_{\theta \to \theta_0} \frac{f(x, \theta) - f(x, \theta_0)}{(\theta - \theta_0)} = \frac{\partial}{\partial \theta} f(x, \theta)|_{\theta = \theta_0},$$

by which I mean that the indicated limit exists, and there exists a function $g(x, \theta_0)$, $\int_{\mathbb{R}} |g|\, dx < \infty$, such that $|(f(x, \theta_0 + \delta) - f(x, \theta_0))/\delta| \leq g(x, \theta_0)$ uniformly in x for all δ small, then

$$\frac{\partial}{\partial \theta} \left[\int_{\mathbb{R}} f(x, \theta)\, dx \right] |_{\theta = \theta_0} = \int_{\mathbb{R}} \left[\frac{\partial}{\partial \theta} f(x, \theta)|_{\theta = \theta_0} \right] dx.$$

Applications to finding how n'th moments change with parameters.

For summation and differentiation, if $\sum_n h(\theta, n)$ converges for all θ in an interval, and $h(\cdot, n)$ is continuously differentiable $\forall n$, and $\sum_n \frac{\partial}{\partial \theta} h(\theta, n)$ converges uniformly on compact subsets of the interval, then $\frac{\partial}{\partial \theta} \sum_n h(\theta, n) = \sum_n \frac{\partial}{\partial \theta} h(\theta, n)$. Application to find $E\, X$ for geometric.

5.3. **Problems.**

Problem 5.1. *Cassella and Berger, Chapter 1, #9, 12, 25, 33, 36, 38, 39, 52, 53, 55.*

Problem 5.2. *Cassella and Berger, Chapter 2, #14, 15, 18, 28, 33, 39.*

Problem 5.3 (Legal theory and Bayes' Theorem)**.** *Suppose that your prior probability that the new fellow in town is a Werewolf (the event B) is $P(B) = 0.001$ (this means that you watch too much late night TV). Now let us suppose that the probability that someone's*

eyebrows grow straight across (the event A) is $P(A) = 0.0025 = \frac{1}{400}$. That is, ahead of time, you would have thought it rather unlikely that the new fellow in town has eyebrows that grow straight across. On the other hand, your exiled Hungarian step-granmother who grew up on the cold slopes of the Caucasus mountains (facing the Black Sea) has told you that 99 times out of a 100, a werewolf's human form will have eyebrows that grow straight across. She's your grandmother, so you believe everything she says. Thus, there is this moderately rare condition, A, which is almost always true if B is true.

1. *What is $P(B|A)$? That is, if the new fellow's eyebrows grow straight across, what is the probability that he's a Werewolf? [If he's a clever Werewolf, he'll shave the space between the eyebrows, but let's ignore this possibility.]*
2. *What aspect(s) should an event E have to be good evidence that the fellow is* **not** *a werewolf? Should it be common? Rare? Should its intersections with B and B^c be large? Small?*

Problem 5.4 (Planting and Leibniz's Rule)**.** *The monsoons will come at some random time T in the next month, $T \in [0, 1]$. A farmer must pre-commit to a planting time, a. As a function of the action, a, and the realized value of T, t, the harvest will be*

$$h(a,t) = \begin{cases} K - r|a-t| & \text{if } a \leq t \\ K - s|a-t| & \text{if } a > t \end{cases}$$

where $K > r, s > 0$. The random arrival time of the monsoon has cumulative distribution function $F(\cdot)$, and $f(x) = F'(x)$ is its strictly positive probability density function.

1. *Graph the function $h(a, \cdot)$ for two different values of a.*
2. *Suppose that the farmer wants to maximize $V(a) = E\, h(a, T)$. Show that the solution to the problem $\max_{a\in[0,1]} V(a)$ is the point $a^* = a^*(r, s)$ satisfying $F(a^*) = r/(r + s)$. (Leibniz' rule will be useful here. Be sure to check second derivatives.)*
3. *Find $\partial a^*/\partial r$ and $\partial a^*/\partial s$. Tell me whether these derivatives are positive or negative and explain the intuition for your answers.*
4. *Now suppose that the last n monsoon times came at the iid times $T_1, \ldots, T_n$. Find an estimator, $\hat{a}_n^*$, of a^*, with the property that $MSE(\hat{a}_n^*) \to 0$ as $n \uparrow \infty$.*

Problem 5.5 (Some hazard rate models)**.** *When $T > 0$ is a random variable with a density giving the random length of life of something, or the random waiting time till the next event, then the* ***hazard rate at*** *t is defined by*

$$h_T(t) = \lim_{\delta \downarrow 0} \frac{P(t \leq T < t + \delta | T \geq t)}{\delta}.$$

This is the proportional rate of change of the probability of surviving another instant given that survival to t has happened, $h_T(t) = -\frac{d}{dt} \ln(1 - F_T(t))$.

Suppose that $X \sim$ *exponential*(β)*, i.e.* $X > 0$ *is a random variable with the property that* $P(X > t) = e^{-t/\beta}$*. Define* $g_\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ *by* $g(r) = r^{1/\gamma}$*.*

1. $Y = g_\gamma(X)$ *has a Weibull* (γ, β) *distribution.*
2. *If* $T \sim$ *exponential*(β)*,* $h_T(t) \equiv 1/\beta$*.* [*Note the memorylessness of this.*]
3. *If* $T \sim$ *Weibull*(γ, β)*,* $h_T(t) = \frac{\gamma}{\beta} t^{\gamma-1}$*.* [*You should vary* γ *above and below* 1*, notice the graphs of* $g(x) = x^\gamma$ *and see why we have the increasing or decreasing hazard rate properties here.*]
4. *If* $T \sim$ *logistic*(β)*,* $h_T(t) = \frac{1}{\beta} F_T(t)$ *where* $F_T(t) = (1 - e^{-(t-\mu)/\beta})^{-1}$*.*

6. Some Continuous Distributions

This section is mostly about getting used to doing calculations and using some of the common distributions. We'll look at discrete and cts distributions, scale and shift them (location and scale families). A particularly important class is the class of exponential distributions.

6.1. **Uniform distributions, $U[\theta_1, \theta_2]$.** The **probability integral transformation** shows that every random variable is a transformation of the uniform distribution. First the uniform distribution, then a specific example, then the general construction.

A random variable X has the uniform distribution on $[0, 1]$, written $X \sim U[0, 1]$, if it has density

$$f_X(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

The cdf is

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

[Note that you can give the density or the cdf to specify a continuous random variable, where by "continuous rv" I mean one with its cdf being the integral of its derivative.]

Suppose that $X \sim U[0, 1]$ and $Y = aX + b$. Then $Y \sim U[b, b + a]$. Give the density and cdf. Going in the reverse direction, if $Y \sim U[\theta_1, \theta_2]$, then $X = \frac{Y-\theta_1}{\theta_2-\theta_1} \sim U[0, 1]$.

Now, take the transformation $Y = g(X)$, where $g(x) = -\log x$. $g(\cdot)$ is a monotonically decreasing function on the interval $(0, 1)$, check the derivative, $g((0, 1)) = (0, +\infty)$. To get the cdf of Y,

$$F_Y(y) = P(Y \leq y) = P(-\log x \leq y) = P(\log x \geq -y) = P(x \geq e^{-y}) = 1 - F_X(e^{-y}) = 1 - e^{-y}.$$

This gives you the negative exponential rv's as monotonic transformations of $U[0, 1]$.

Given a cdf, $F_Y(x)$, it is possible to express the random variable Y as the weakly monotonic transformation of X where $X \sim U[0, 1]$. Show how. This is a method used to generate rv's for simulation.

Suppose that $\theta \in \Theta = [0, \infty)$ is the unknown size of the largest fish in a given body of water. Suppose that $P_\theta = U[0, \theta]$, so that if θ is true, then the data, the size of the n fish you've caught, are independent and uniformly distributed over the interval $[0, \theta]$. Here are two possible estimators:

$$\hat{\theta}_1(X_1, \ldots, X_n) = \max\{X_i : i = 1, \ldots, n\},$$

$$\hat{\theta}_2(X_1, \dots, X_n) = 2 \cdot \frac{1}{n} \sum_{i=1}^{n} X_i.$$

We know that $\hat{\theta}_2$ is not biased, that is, $E_\theta \hat{\theta}_2 = \theta$. We also know that $\hat{\theta}_1$ is biased downwards, that is $E_\theta \hat{\theta}_1 < \theta$. Being unbiased, we can see that a substantial part of the time, $\hat{\theta}_2$ will be a really stupid estimator, that is, we'll have $\hat{\theta}_2 < \hat{\theta}_1$, and since we know that $\hat{\theta}_1 < \theta$, this is really not sensible. This suggests the estimator $\hat{\theta}_3 := \max\{\hat{\theta}_1, \hat{\theta}_2\}$, which is less biased than $\hat{\theta}_1$ and less regularly stupid than $\hat{\theta}_2$. Yet another approach is Bayesian, suppose that your prior distribution is that $\theta \sim$exponential(2), write out the likelihood, look at the MLE.

The moral of the story, for later purposes, is that there are many estimators and we need ways to choose between them.

6.2. **The normal or Gaussian family of distributions, $N(\mu, \sigma^2)$.** The random variable Z has the **standard normal distribution**, aka the **standardized Gaussian**, if it has density

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \ -\infty < z < \infty.$$

Important: From now on, we will reserve Z to mean a random variable with this distribution. It is written $Z \sim N(0, 1)$.

The constant $\frac{1}{\sqrt{2\pi}}$ is there to make the density integrate to 1, that is, $\int_{-\infty}^{+\infty} e^{-z^2/2}\, dz = \sqrt{2\pi}$, a result that comes from changing to polar coordinates and doing a rotation integral.

No-one knows a closed form formula for the cdf of the Gaussian, that is

$$F_Z(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

is the best we can do. This cdf turns out to be so useful that its numerical values have been calculated to very high degrees of precision and the results have been tabulated. These are included in all statistical software and many spread sheets.

z^2 is symmetric about 0, so that $e^{-z^2/2}$ is symmetric. This implies that $E\, Z = 0$. That was the easy way to get the result, now we're going to do it by change of variable, a technique that is useful enough that simple reviews are worthwhile.

$$E\, Z = \int_{-\infty}^{+\infty} z e^{-z^2/2}\, dz = \int_{-\infty}^{0} z e^{-z^2/2}\, dz + \int_{0}^{+\infty} z e^{-z^2/2}\, dz.$$

Using the change of variable $x = -z$ so that $dz = -dx$, and noting that $(-z)^2 = z^2$,

$$\int_{-\infty}^{0} z e^{-z^2/2}\, dz = \int_{0}^{+\infty} -x e^{-x^2/2}\, dx = -\int_{0}^{+\infty} x e^{-x^2/2}\, dx.$$

Combining,

$$E\,Z = \int_{-\infty}^{0} ze^{-z^2/2}\,dz + \int_{0}^{+\infty} ze^{-z^2/2}\,dz = -\int_{0}^{+\infty} ze^{-z^2/2}\,dz + \int_{-\infty}^{0} ze^{-z^2/2}\,dz = 0.$$

If $Z \sim N(0,1)$ and $Y = \sigma Z$, then we write $Y \sim N(0,\sigma^2)$, and $E\,Y = 0$ and $\mathrm{Var}(Y) = \sigma^2$. Since Z is symmetric about 0, the sign of σ does not matter, and by convention, $\sigma > 0$. We need to get the density of Y. We'll get an expression for the cdf, and use the chain rule to get the density.

$$F_Y(a) = P(Y \le a) = P(\sigma Z \le a) = P(Z \le a/\sigma) = F_Z(a/\sigma).$$

Therefore,

$$f_Y(a) = F_Y'(a) = \frac{d}{da}F_Z(a/\sigma) = f_Z(a/\sigma)\cdot\frac{1}{\sigma} = \frac{1}{\sigma}\frac{1}{\sqrt{2\pi}}\,e^{-(a/\sigma)^2/2} = \frac{1}{\sigma\sqrt{2\pi}}\,e^{-a^2/2\sigma^2}.$$

If $Y \sim N(0,\sigma^2)$ and $X = Y+\mu$, then we write $X \sim N(\mu,\sigma^2)$, and $E\,X = \mu$, $\mathrm{Var}(X) = \sigma^2$. To get the density of X, we note that

$$F_X(a) = P(X \le a) = P(Y+\mu \le a) = P(Y \le (a-\mu)).$$

Applying the chain rule again, this yields

$$f_X(a) = f_Y(a-\mu) = \frac{1}{\sigma\sqrt{2\pi}}\,e^{-(a-\mu)^2/2\sigma^2}.$$

Going in the other direction, if $X \sim N(\mu,\sigma^2)$, then the standardized version of X is the rv $X_s = \frac{X-\mu}{\sigma}$, and $X_s \sim N(0,1)$. Shifting and scaling are all that are at work in all of this. Go through a couple of problems on standardizing and then reading from the normal tables.

6.3. **A useful device. Claim**: (Thm. 3.5.1 in text) If $f(x)$ is a pdf, then for all μ and all $\sigma > 0$, $g(x|\mu,\sigma) = \frac{1}{\sigma}f(\frac{x-\mu}{\sigma})$ is a pdf. We can see why, this is shifting and scaling an rv.

6.4. **The gamma family, $\Gamma(\alpha,\beta)$.** This is a family of strictly positive rvs. For $y > 0$, $y^{\alpha-1}e^{-y} > 0$, and it should be pretty easy to believe that for any $\alpha > 0$, the integral $\int_0^\infty y^{\alpha-1}e^{-y}\,dy$ is finite. Therefore, if we define the function $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}\,dy$, then for every $\alpha > 0$, the following is a density,

$$f(y) = \frac{1}{\Gamma(\alpha)}y^{\alpha-1}e^{-y}, \quad y > 0.$$

Such a rv is written $Y \sim \Gamma(\alpha,1)$, and the role of the 1 will become clear soon.

Look at a couple values of α, e.g. $\alpha = 1,2,4$, we get a sense of what this class of distributions is about.

You should check that $\Gamma(1) = 1$. Integration by parts gives $\Gamma(\alpha+1) = \alpha \cdot \Gamma(\alpha)$ for $\alpha > 0$. Therefore, $\Gamma(2) = (2-1)\cdot 1 = (2-1)!$, $\Gamma(3) = (3-1)\cdot(2-1)! = (3-1)!$, $\Gamma(4) = (4-1)\cdot(3-1)! = (4-1)!$, $\dots$, $\Gamma(n) = (n-1)!$.

Review of integration by parts:

$$\int_a^b f dg = fg|_a^b - \int_a^b g df,$$

this comes from the product rule, $dfg = fdg + gdf$ so that $fdg = dfg - gdf$.

$$\Gamma(\alpha+1) = \int_0^\infty y^\alpha e^{-y} dy,$$

set $f = y^\alpha$, $dg = e^{-y}dy$ so that $df = \alpha y^{\alpha-1}$ and $g = -e^{-y}$, and we have

$$\int_0^\infty y^\alpha e^{-y} dy = -y^\alpha e^{-y}|_0^\infty - \int_0^\infty \alpha y^{\alpha-1}(-e^{-y}) dy = (0-0) + \alpha \int_0^\infty y^{\alpha-1} e^{-y} dy = \alpha\Gamma(\alpha-1).$$

It may not come as a surprised that when α is not an integer, we do not have any closed form expression for $\int_c^d \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y}\, dy$, so, we have tables and numerical integration programs to find it for us.

If $Y \sim \Gamma(\alpha, 1)$, then $EY = \mathrm{Var}(Y) = \alpha$. This is not at all obvious until you've fooled with the integrals a bit.

$$\int_0^\infty y \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y}\, dy.$$

Now, for all $\alpha > 0$, $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\, dy$. Therefore,

$$EY = \frac{1}{\Gamma(\alpha)} \int_0^\infty y y^{\alpha-1} e^{-y}\, dy = \frac{1}{\Gamma(\alpha)} \int_0^\infty y^{(\alpha+1)-1} e^{-y}\, dy = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} = \alpha.$$

The same logic (more or less) tells us that $EY^2 = \alpha(\alpha+1)$, so that $\mathrm{Var}(Y) = \alpha(\alpha+1) - \alpha^2 = \alpha$.

The Γ class of distributions can be scaled, but not shifted. This last is mostly for convenience, we want them all to be distributed on the interval $[0, \infty)$.

If $Y \sim \Gamma(\alpha, 1)$, and $X = \beta \cdot Y$, $\beta > 0$, then we can use the same technique we had above to find the density of X.

$$F_X(x) = P(X \le x) = P(\beta Y \le x) = P(Y \le x/\beta) = F_Y(x/\beta) = \int_0^{x/\beta} \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y}\, dy.$$

Therefore,

$$f_X(x) = \frac{d}{dx} F_Y(x/\beta) = f_Y(x/\beta) \cdot \frac{1}{\beta} = \frac{1}{\beta} \frac{1}{\Gamma(\alpha)} (x/\beta)^{\alpha-1} e^{-(x/\beta)}.$$

After re-organizing the β terms, we get

$$f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-(x/\beta)},$$

and we write this as $X \sim \Gamma(\alpha, \beta)$. This gives a large class of distributions connected to any number of random phenomena.

From the usual rules, if $X \sim \Gamma(\alpha, \beta)$, then $E\,X = \alpha\beta$ and $\mathrm{Var}(X) = \alpha\beta^2$.

6.5. **Special cases of $\Gamma(\alpha, \beta)$ distributions.** There are some special $\Gamma(\alpha, \beta)$ distributions, ones that have their own special names and uses.

6.5.1. *Waiting times.* If $Y \sim \Gamma(1, \beta)$, we say that Y has an **exponential distribution with parameter** β. We've seen this as the waiting time for the first Poisson arrival with an arrival rate $\lambda = 1/\beta$. From above, $E\,Y = \beta$, and $\mathrm{Var}(Y) = \beta^2$.

Example: Suppose that $Y \sim \Gamma(1, 100)$ and $X = \min\{Y, 200\}$, find the cdf and the expectation of X.

Along this line, it turns out that all of the Poisson waiting time distributions are contained in the Γ family. If $Y \sim \Gamma(\alpha, 1)$, α an integer, then for any $t > 0$,

$$P(Y > t) = \int_t^\infty \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y}\, dy = \sum_{n=0}^{\alpha-1} \frac{t^n e^{-t}}{n!} = P(\text{Poisson}(t) \le \alpha - 1).$$

We already know this result for $\alpha = 1$. Getting the rest of the α is a good exercise in applying integration by parts.

6.5.2. *Squares of standard normals.* If $Y \sim \Gamma(v/2, 2)$, v an integer, then Y has a $\chi^2_{(v)}$ **distribution**, read as a **chi squared distribution with v degrees of freedom**. Reading off directly, we have

$$f_Y(y) = \frac{1}{\Gamma(\frac{v}{2})} x^{\frac{v}{2}-1} e^{-x}.$$

Directly checking, you can see that if $Z \sim N(0, 1)$, i.e. has density $(2\pi)^{-\frac{1}{2}} e^{\frac{x^2}{2}}$, the density of Z^2 is $f_Y(y) = \frac{1}{\Gamma(\frac{1}{2})} x^{\frac{1}{2}-1} e^{-x}$.

Check that the moment generating function of a $X \sim \Gamma(\alpha, \beta)$ is $M_X(t) = E\,e^{tX} = \left(\frac{1}{1-\beta t}\right)^\alpha$. The basic result is that mgf's identify random variables that have them, find $M_{X_1+\cdots+X_v}(t)$ when the X_i are iid χ^2_1's.

6.6. **Cauchy random variables.** From calculus we know that $d\arctan(t)/dt = (1+t^2)^{-1}$, so that

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}$$

is a density. It is the standard Cauchy density. By shifting location, we get the family

$$f(x|\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}.$$

The MLE estimator of θ is weird, check it's behavior. Some of the weirdness comes from the fact that if $X_1, \ldots, X_n$ are iid Cauchy(0), then $\frac{1}{n}\sum_{i\leq n} X_i$ is Cauchy(0). We get this easily after we do that harder calculation that $X \sim$Cauchy(0) implies that $\varphi_X(t) = E\, e^{itX} = e^{-|t|}$ and doing the calculation for $\varphi_{\frac{1}{n}\sum_{i\leq n} X_i}(t)$.

6.7. **Exponential Families.** A class of pdf's of the form

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta})\exp\left(\sum_{i\leq I} w_i(\boldsymbol{\theta})t_i(x)\right)$$

is called an **exponential family**. Implicit in this notation is that $c(\cdot)$ and the $w_i(\cdot)$ do NOT depend on x, and that $h(\cdot)$ and the $t_i(\cdot)$ do NOT depend on θ.

Look at the MLE's.

E.g. binomial(n,p), $0<p<1$, is

$$f(x|p) = {}_nC_x(1-p)^n \exp\left(\log\left(\frac{p}{1-p}\right)x\right),$$

so that $h(x) = {}_nC_x 1_{\{0,1,\ldots,n\}}(x)$, $c(p) = (1-p)^n$, $I=1$, $w_1 = \log(\frac{p}{1-p})$, and $t_1(x) = x$.

E.g. normal$(\mu,\sigma^)$ is an exponential family.

Not all families are exponential, a basic problem arises when the support depends on $\boldsymbol{\theta}$. If the mapping $\boldsymbol{\theta} \mapsto (w_i(\boldsymbol{\theta}))_{i\leq I}$ is one-to-one and invertible, we can replace the $\boldsymbol{\theta}$'s by η_i's in the reparametrization

$$f(x|\eta) = h(x)c^*(\eta)\exp\left(\sum_{i\leq I}\eta_i t_i(x)\right),$$

and the set of η_i's that make this a density is the so-called **natural parameter space** for the class of densities. It's convex, which is nice.

6.8. **Some (in)equalities.** Tchebyshev and it's higher moment versions. The 3σ rule. $P(|Z|\geq t) \leq \sqrt{2/\pi}\cdot e^{-t^2/2}/t$, from

$$P(Z\geq t) = 1/\sqrt{2\pi}\int_t^\infty e^{-x^2/2}dx \leq 1/\sqrt{2\pi}\int_t^\infty \frac{x}{t}e^{-x^2/2}dx = 1/\sqrt{2\pi}e^{-t^2/2}/t.$$

If $X\sim N(\theta,\sigma^2)$, g continuously differentiable and $E|g'(X)|<\infty$, then $E[g(X)(X-\theta)] = \sigma^2 E\, g'(X)$. E.g. $g(x) = ax+b$, or $g(x) = x^2$ (which gives $E\, X^3 = 3\theta\sigma^2+\theta^3$. [Integrate by parts with $u=g$, $dv = (x-\theta)e^{-(x-\theta)^2/2\sigma^2}$.]

6.9. **Problems.**

Problem 6.1. *Casella & Berger, 3.1.*

Problem 6.2. *Casella & Berger, 3.2.*

Problem 6.3. *Casella & Berger, 3.3.*

Problem 6.4. *Casella & Berger, 3.4.*

Problem 6.5. *Casella & Berger, 3.6.*

Problem 6.6. *Casella & Berger, 3.7.*

Problem 6.7. *Casella & Berger, 3.9.*

Problem 6.8. *Casella & Berger, 3.15.*

Problem 6.9. *Casella & Berger, 3.16.*

Problem 6.10. *Casella & Berger, 3.20.*

Problem 6.11. *Casella & Berger, 3.24.*

Problem 6.12. *Casella & Berger, 3.28.*

Problem 6.13. *Casella & Berger, 3.29.*

Problem 6.14. *Casella & Berger, 3.31 and 3.32.*

Problem 6.15. *Casella & Berger, 3.41.*

7. Random Vectors, Conditional Expectations, Independence

We now turn to modeling the simultaneous randomness of many quantities.

7.1. **Dependence, conditional probabilities and expectations.** The starting point is dependence, $P(Y \in A|X \in B)$, down to $P(Y|X = x)$. In the discrete case, use Bayes' Law. E.g. U is the random number of rolls of two fair die until the first sum of 7, V until the second, give joint pdf, $P(V = v|U = u)$, $E(V|U = u)$, $P(U = u|V = v)$, $E(U|V = v)$, etc.

In the continuous case, $f_Y(y|x)dy = \frac{f_{X,Y}(x,y)dxdy}{f_X(x)dx}$, cancelling terms, $f_Y(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$. E.g. the two light bulb story, both exponential parameter 1, $f_{X,Y}(x,y) = 1_A(x,y)e^{-y}$, $A = \{(u,v) \in \mathbb{R}^2 : v > u > 0\}$. $E(Y|X = x) := \int y f_Y(y|x)\,dy$, $\text{Var}(Y|X = x) := \int (y - E(Y|X = x))^2 f_Y(y|x)\,dy$, or $\text{Var}(Y|X = x) = E(Y^2|x) - (E(Y|x))^2$. The main applications of conditional variance models are in financial economics, summarize some volatility studies.

A very subtle concept is $E(Y|X)$, since X is a rv ... . It is a random variable that depends on X. The law of iterated expectations is $EY = E\,E(Y|X)$. It must be true in the discrete case (show why), and in the continuous case too (show why).

The following is the justification for most of regression analysis, linear and non-linear too. Amongst the functions $g(x)$ depending only on x, the one that solves the problem

$$\min_{g(x)} E(Y - g(X))^2$$

is the function $g(x) = E(Y|X = x)$. The function $E(Y|X)$ is called the **regression of Y on X**. Do this first by conditioning on $X = x$ and looking point by point, then more generally. The mapping from Y to $E(Y|X)$ is a projection.

7.2. **Projections.** Let M be a linear subspace of $\mathbb{R}^n$. $N = M^\perp$ is the orthogonal complement of M. Every $x \in \mathbb{R}^n$ has a unique representation as $m + n$, $m \in M$, $n \in N$. $P : \mathbb{R}^n \to M$ is a projection if $Px = m$ where $x = m + n$, $n \perp M$, is the unique representation of x. Note that P is linear, that $P(\mathbb{R}^n) = M$, $P^2 = P$, and $P_{|M} = I_{|M}$.

Suppose that we observe n values of y, arranged in an $n \times 1$ vector Y, and n values of x_i, $i = 1, \ldots, k$, arranged in an $n \times k$ matrix X. We assume that $n > k$ and that the columns of X are independent. We are after the β that makes

$$f(\beta) := (Y - X\beta)'(Y - X\beta) = e'_\beta e_\beta$$

as small as possible where $e_\beta = Y - X\beta$. Let M be the column span of X. Do the algebra, get the classic formula, and the projection and residual matrixes. Show that e^* is perpendicular to M, and that you could just look for β to make that happen.

There is a connection to the normal distribution. Suppose that $Y_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$, find the MLE of μ, σ^2. Now suppose that $Y_i \sim N(X_i\beta, \sigma^2)$ are independent $i = 1, \ldots n$ for some β, σ^2. Write the likelihood function, solve for β.

One kind of hypothesis of interest is (say) $\beta_k = 0$. Solve the minimization restricted problem for β_r and compare $\|e_{\beta_r}\|$ to $\|e_\beta\|$. We can tell which of these is going to be bigger, it's projection onto $M' \subset M$. If these are too far apart, we have evidence that the hypothesis should be rejected. The actual distributions involved can be complicated, but not too miserable.

Now let $M \subset L^2$ be the set of functions of X such that $E\, g(X)^2 < \infty$. Define $P : Y \to M$ by $P(Y) = E(Y|X)$. P is a projection.

7.3. **Causality and conditional probability.** There is a strong distinction between causality and conditional probability. Noting that $P(A|B) > P(A)$ does NOT mean that B causes A. It only means that knowing that B is true makes us infer that A is more likely.

The conditional probability of the death of a patient rises with the amount of praying done in their room, but this does not lead us to conclude that prayer causes death. $P(A|B) > P(A)$ means that when we observe B, we are more likely to observe A, not that B causes A.

If B is the event that there was a nightlight in your room as a child and A is the event that you need corrective lenses when you are an adult, then $P(A|B) > P(A)$. This is NOT causality at work.

In a similar fashion, $\psi(x) := E(Y|x)$ and $\psi'(x) > 0$ is not an indication that higher values of X cause higher values of Y, rather, it is an indication that higher values of X are observed with higher average values of Y. It is SO tempting to assert causality from such a result. You've got to use that resource that Twain said was not common . . . .

Historically, if B is the event that interest rates are above average and A is the event that GNP growth is above average, $P(A|B) > P(A)$. Causality? Probably not. There are many other examples.[2]

[2]The following is from `http://my.webmd.com/content/article/1836.50533`, a recent article about sleep and longevity that also appeared in the local newspaper. It also included the figures that the average amount of sleep/night is 7 hours, that, on average, those who slept 7 hours/night had the lowest death rate (probability of death in any given year), that sleeping 8 hours is associated with a probability of death that is 12% higher than the lowest, 9 hours is associated with a probability of death that is 17% higher than the lowest, 10 hours is associated with a probability of death that is 34% higher than the lowest.

> Kripke and co-workers analyzed data from an American Cancer Society study conducted between 1982 and 1988. The study gathered information on people's sleep habits and health, and then followed them for six years. Study participants ranged in age from 30 to 102 years, with an average starting age of 57 years for women and 58 years for men.
> Death risk increased for those who go too little sleep, too, but the numbers are smaller. The risk of death went up 8% for those who slept six hours, 11% for those who slept five hours, and 17% for those who slept only four hours a night.
> While this increased risk is statistically significant, it doesn't translate into much of a risk for an individual person. The study's main finding, Kripke says, is that sleeping less than eight hours isn't bad for you. In fact, eight hours' sleep can no longer be considered normal.

7.4. **Independence, sums of independent rv's.** Review $X \perp\!\!\!\perp Y$, for all A, B, $P([X \in A] \cap [Y \in B]) = P([X \in A]) \cdot P([Y \in B])$. In particular, for all $g(x)$ depending only on x and all $h(y)$ depending only on y, $g(X) \perp\!\!\!\perp h(Y)$.

Look at product support sets, product cdf's, product pdf's. Note that $\frac{d}{dx}P(Y \in A|X = x) = 0$ and therefore $\frac{d}{dx}E(YA|X = x) = 0$ when $X \perp\!\!\!\perp Y$.

A basic result is

Claim: If $X \perp\!\!\!\perp Y$, then for all $g(x)$ depending only on x and all $h(y)$ depending only on y, $E\, g(X)h(Y) = (E\, g(X)) \cdot (E\, h(Y))$. In particular, $E\, XY = (E\, X) \cdot (E\, Y)$.

Example where X is not independent of Y and this fails — $X \sim U(-1, +1)$, $Y = -X$, $g(x) = x$, $h(y) = y$, $E\, g(X)h(Y) = E\,(-X^2) = -\frac{1}{3}$, while $E\, g(X) = 0 = E\, h(Y)$. Give other examples too.

Argument for Claim: $E\, g(X)h(Y) = \int\int g(x)h(y)f_X(x)f_Y(y)\, dxdy = (E\, g(X)) \cdot (E\, h(Y))$.

Note that setting $g(x) = 1_A(x)$ and $h(y) = 1_B(y)$ gives back our definition of independence.

Claim: If $X \perp\!\!\!\perp Y$ and $Z := X + Y$, then $M_Z(t) = M_X(t) \cdot M_Y(t)$.

If $X \sim n(\mu, \sigma^2)$, then $M_X(t) = e^{\mu t + \sigma^2 t^2/2}$, so that both sums and variances add for Normal distributions.

If $X \perp\!\!\!\perp Y$, $X \sim$ Poisson(θ), $Y \sim$ Poisson(λ), then $X + Y \sim$ Poisson$(\theta + \lambda)$. This can be had by brute calculation (as in the text), or by noting that $M_X(t) = e^{\theta(e^t - 1)}$ when $X \sim$ Poisson(θ), and being watchful.

The Normal and the Poisson are **infinitely divisible**, they are the two major types of rv's to come out of the big CLT.

7.5. **Covariance and correlation.** Given two rv's X, Y with means μ_X, μ_Y and variances σ_X^2, σ_Y^2, we are interested in a numerical measure of the relatedness of X and Y. It comes

---

So why does it feel good to sleep in? Oversleeping may be a lot like overeating, suggests Jim Horne, PhD, director of the sleep research center at Loughborough University, England.

"As we can eat more food than we require and drink more fluids than we require, or drink beer, or eat foods we don't need, we may sleep more than we require," Horne tells WebMD. "There is an optionality about it. The amount of sleep we require is what we need not to be sleepy in the daytime."

The Kripke study also shows that people who say they have insomnia aren't necessarily in bad health. But those who often take sleeping pills have an increased risk of death. Frequent use of sleeping pills increased the risk of death by 25%.

"The risk of taking a sleeping pill every night is equivalent to sleeping three or 10 (*sic*) hours," Kripke says. "It is a substantial risk factor. We cannot say it causes these deaths or that this risk applies to newer medicines. But lacking evidence for safety, the wisest choice is to be cautious in their use."

Bliwise says it's always a good idea to be cautious about using sleeping pills. However, he sees no real problem in the proper use of these drugs from time to time. "There is no data that intermittent use of a short acting [prescription sleeping pill] is necessarily harmful," he says.

through the function $g(x, y) = (x - \mu_X)(y - \mu_Y)$, which looks like a Cobb-Douglas utility function with the origin moved to (μ_X, μ_Y). $g > 0$ to the NE and SW of the shifted origin and $g < 0$ to the NW and SE of the shifted origin. When g is, on average, positive (negative), we have evidence of a positive (negative) relation between X and Y.

Definition: $\sigma_{X,Y} = \text{Cov}(X, Y) := E\, g(X, Y) = E\,(X - \mu_X)(Y - \mu_Y)$.

Definition: $\rho_{X,Y} = \text{corr}(X, Y) := \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$.

Note that $\sigma_{X,Y} = \text{Cov}(X, X) = \sigma_X^2 \geq 0$, and is strictly positive as long as X is not degenerate. Note also that $\sigma_{X,Y} = \sigma_{Y,X}$ and $\rho_{X,Y} = \rho_{Y,X}$. Show that $\text{Cov}(X, Y) = E\,XY - \mu_X \mu_Y$. We use correlation because it is a unitless measure of the relation between X and Y, show why. Another note: if $P(X = c) = 1$, then $\text{Cov}(X, Y) = E\,(c - c)(Y - \mu_Y) = 0$ implying that $\rho_{X,Y} = 0$. There are more interesting ways to get 0 correlation, the one we see most often is independence.

Claim: If $X \perp\!\!\!\perp Y$, then $\text{Cov}(X, Y) = 0$, but $[\text{Cov}(X, Y) = 0] \not\Rightarrow [X \perp\!\!\!\perp Y]$.

Show why, an easy continuous counter-example is (X, Y) uniformly distributed over the disk of radius 1 centered at 0, a simple discrete example is the uniform distribution over the 5 points $(-1, -1)$, $(-1, +1)$, $(+1, -1)$, $(+1, +1)$, and $(0, 0)$.

7.6. **Bivariate normals.** If X_1, X_2 have the joint density

$$f(x_1, x_2) = k e^{-\frac{1}{2}(x-\mu)' \Sigma (x-\mu)},$$

where $k = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$, Σ the variance-covariance matrix, μ the mean vector, then we have a bivariate normal distribution. Note that the part of denominator of k that is not about 2π is the square root of the determinant of Σ.

If Z_1, Z_2 are independent $N(0, 1)$'s, and X and Y are different affine combinations of the Z's, then X and Y are bivariate normals with the appropriate parameters.

7.7. **A pair of discrete, portfolio management examples.** Suppose that X and Y are the random variables describing the rate of return on two different investments. We're going to consider two possible joint distributions of X and Y, one with negative and one with positive correlation. To keep things simple, the two joint distributions have the same marginal distributions.

A

1.10	$\frac{1}{3}$	$\frac{1}{6}$
1.04	$\frac{1}{6}$	$\frac{1}{3}$
$Y\uparrow/X\rightarrow$	1.00	1.22

B

1.10	$\frac{1}{6}$	$\frac{1}{3}$
1.04	$\frac{1}{3}$	$\frac{1}{6}$
$Y\uparrow/X\rightarrow$	1.00	1.22

The marginal distributions are, in all four cases, $(\frac{1}{2}, \frac{1}{2})$. Eyeballing the two cases, A has negative correlation and B positive between X and Y.

1. $\mu_X = 1.11$, $\sigma_X^2 = 1.2442 - (1.11)^2 = 0.0121$ so that $\sigma_X = 0.11$.
2. $\mu_Y = 1.07$, $\sigma_Y^2 = 1.1458 - (1.07)^2 = 0.0009$ so that $\sigma_Y = 0.03$.

So, investing in Y gives a lower rate of return but also has a lower variance. The essential portfolio selection problem is what proportion of one's wealth to hold in which assets. Intuitively, the answer should depend on one's attitude toward the risk-return tradeoff, which we'll not study in this class, and the covariance of X and Y, which we will study.

1. In case A, $E\,XY = 1.1866$ so that $\text{Cov}(X,Y) = 1.1866 - (1.11)(1.07) = 1.1866 - 1.1877 = -0.0011$ and $\text{corr}(X,Y) = \frac{-0.0011}{0.11 \cdot 0.03} = -\frac{1}{3}$.
2. In case B, $E\,XY = 1.1888$ so that $\text{Cov}(X,Y) = 1.1888 - (1.11)(1.07) = 1.1888 - 1.1877 = 0.0011$, and $\text{corr}(X,Y) = \frac{0.0011}{0.11 \cdot 0.03} = \frac{1}{3}$.

Claim: $\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X,Y)$.

To see why, start with $\text{Var}(aX+bY) = E\left[(aX+bY) - (a\mu_X + b\mu_Y)\right]^2 = E\left[(a(X - mu_X) + b(Y - b\mu_Y)\right]^2$ and do the obvious algebra.

We're particularly interested in the case where $0 \leq a \leq 1$ and $a + b = 1$. Here we think of a as the share of the portfolio in X and b as the share in Y.

The variance of a portfolio, as a function of a is

$$f(a) = \text{Var}(aX + (1-a)Y) = a^2\text{Var}(X) + (1-a)^2\text{Var}(Y) + 2a(1-a)\text{Cov}(X,Y).$$

This can be re-written as

$$f(a) = a^2[\sigma_X^2 + \sigma_Y^2 - 2\sigma_{X,Y}] + 2a[\sigma_{X,Y} - \sigma_Y^2] + \sigma_Y^2.$$

The term multiplying a^2 is strictly positive (being the variance of $X - Y$). Therefore, this is a parabola opening upwards. $f(0) = \sigma_Y^2$ and $f(1) = \sigma_X^2$, and the only remaining question is where the parabola reaches its minimum.

It's minimum happens when $f'(a) = 0$, that is, when

$$2a[\sigma_X^2 + \sigma_Y^2 - 2\sigma_{X,Y}] + 2[\sigma_{X,Y} - \sigma_Y^2] = 0,$$

that is, when $a = a^*$,

$$a^* = \frac{\sigma_Y^2 - \sigma_{X,Y}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{X,Y}}.$$

Some special cases:

1. When $\sigma_X = \sigma_Y$, $a^* = \frac{1}{2}$.
2. $f'(0) > 0$ gives $a^* = 0$, this happens when $\sigma_{X,Y} - \sigma_Y^2 \geq 0$, that is, when

$$\rho_{X,Y} \geq \frac{\sigma_Y}{\sigma_X}.$$

This NEVER happens if $\rho_{X,Y} < 0$, and there is a good intuition, negative correlation means that bad shocks to one stock are (somewhat) likely to be offset by good shocks to the other, and vv. $a^* = 0$ can only happen when $\rho_{X,Y} > 0$ and σ_Y is (very) small relative to σ_X, that is, Y has much less variance and is positively correlated with the much riskier X.

3. Doing the algebra for $f'(1) < 0$, which gives $a^* = 1$, gives the condition
$$\rho_{X,Y} \geq \frac{\sigma_X}{\sigma_Y},$$
and again, this NEVER happens if $\rho_{X,Y} < 0$, and all that has changed is that X must be the small variance random variable.

Time to go back to the first special case and examine it in terms of the conditions developed in the second two cases.

Claim: $-1 \leq \rho_{X,Y} \leq 1$, and when both σ_X and σ_Y are strictly positive, $\rho_{X,Y} = 1$ iff $Y = aX + b$ for some $a > 0$, while $\rho_{X,Y} = -1$ iff $Y = aX + b$ for some $a < 0$.

If $\sigma_X = \sigma_Y$, then we will never have $\rho_{X,Y} > \sigma_X/\sigma_Y = \sigma_Y/\sigma_X$ because this requires $\rho_{X,Y} > 1$.

The text proves the last claim by using a clever device involving the determinant of the polynomial $h(t) = \text{Var}((X-\mu_X)t+(Y-\mu_Y)) = E\left[(X-\mu_X)t+(Y-\mu_Y)\right]^2 = t^2\sigma_X^2+2t\sigma_{X,Y}+\sigma_Y^2$. Now, $h(t) \geq 0$, but the Fundamental Theorem of Algebra tells us that it has two roots (counting multiplicity) in the complex plane. The discriminant is $(2\sigma_{X,Y})^2 - 4\sigma_X^2\sigma_Y^2$, and this must therefore be less than or equal to 0. Rearrange to get $-1 \leq \rho_{X,Y} \leq 1$. We have $|\rho_{X,Y}| = 1$ iff the discriminant is equal to 0, that is, iff the polynomial $h(t)$ has a single real root, call it t^*. Then we know that $\text{Var}((X - \mu_X)t^* + (Y - \mu_Y)) = 0$, that is, $P[(X - \mu_X)t^* + (Y - \mu_Y) = 0] = 1$. Using the quadratic formula, $t^* = -\sigma_{X,Y}/\sigma_X^2$, which finishes everything up.

Another, even more clever argument runs through the Cauchy-Schwarz inequality: For any a, b, $(a-b)^2 \geq 0$ with equality iff $a = b$. This means that $a^2-2ab+b^2 \geq 0$ with equality iff $a = b$. This in turn means that $a^2 + b^2 \geq 2ab$ or
$$\frac{1}{2}a^2 + \frac{1}{2}b^2 \geq ab \text{ with equality iff } a = b.$$

Now, for each $\omega \in \Omega$, let
$$a(\omega) = \frac{X(\omega)}{\sqrt{E\,X^2}}, \text{ and } b(\omega) = \frac{Y(\omega)}{\sqrt{E\,Y^2}}.$$
Applying the inequality we had above at each ω, we get
$$\frac{1}{2}(a(\omega))^2 + \frac{1}{2}(b(\omega))^2 \geq a(\omega)b(\omega),$$

which is re-written as

$$\frac{1}{2}\frac{X^2(\omega)}{E\,X^2}+\frac{1}{2}\frac{Y^2(\omega)}{E\,Y^2}\geq\frac{XY(\omega)}{\sqrt{E\,X^2}\sqrt{E\,Y^2}}\quad\text{with equality iff } X(\omega)=cY(\omega),$$

where the constant c is equal to $\sqrt{E\,X^2}/\sqrt{E\,Y^2}$, something independent of ω. This last equality holds for all ω, taking expectations, the left hand side is equal to 1, rearranging gives

$$E\,XY\leq\sqrt{E\,X^2}\sqrt{E\,Y^2}.$$

Since $X^2=|X|^2$ and $|XY|=|X|\cdot|Y|$, the same logic gives

$$E\,|XY|\leq\sqrt{E\,X^2}\sqrt{E\,Y^2},$$

known as the Cauchy-Schwarz inequality. Since $-|XY|\leq XY\leq|XY|$, $|E\,XY|\leq E|XY|$. Along with the C-S inequality,

$$|E\,(X-\mu_X)(Y-\mu_Y)|\leq\sqrt{E\,(X-\mu_X)^2}\sqrt{E\,(Y-\mu_Y)^2}.$$

Squaring both sides gives

$$(\mathrm{Cov}(X,Y))^2\leq\sigma_X^2\sigma_Y^2,$$

which tells us that $(\rho(X,Y))^2\leq 1$.

Enough of the purely abstract stuff, some examples will help.

Example: $X\sim U(0,1)$ and $Y=3X$, explicitly calculate $\rho_{X,Y}$.

Example: $X\sim U(0,1)$ and $Y=-3X$, explicitly calculate $\rho_{X,Y}$.

Example: $X\sim U(0,1)$ and $Y=X^2$, explicitly calculate $\rho_{X,Y}$.

Going back to case A, since $\rho_{X,Y}=-\frac{1}{3}<0$, we know that $0<a^*<1$, specifically, it is

$$a^*=\frac{0.0009-(-0.0011)}{0.0121+0.0009-2(-0.0011)}\simeq 0.13,$$

so, to minimize variance in case A, have 13% of your portfolio in X, the rest in Y. Graph a vs. $g(a):=E\,(aX+(1-a)Y)$ and $\sqrt{f(a)}=\sigma_{aX+(1-a)Y}$. Also graph $g(a)$ on the horizontal axis and the corresponding $f(a)$ on the vertical. Note that for mean rates of return less than $1.07+(0.13)\cdot(1.11-1.07)\simeq 1.075$, taking a lower rate of return gives you a higher variance, not generally considered a sensible way to behave. Go through the gradients of a utility function that depends positively on mean and negatively on standard deviation.

Going back to case B, since $\rho_{X,Y}=\frac{1}{3}$, and $\frac{1}{3}>\sigma_Y/\sigma_X=0.27\overline{27}$, we know that $a^*=0$. Give the same two graphs.

7.8. **The matrix formulation.** Let $X=(X_1,\ldots,X_N)^T$, be rv's with means $\mu=(\mu_1,\ldots,\mu_N)^T$, and cross moments $\sigma_{n,m}=\mathrm{Cov}(X_n,X_m)$. Let Σ the the symmetric, $N\times N$ matrix with (n,m)'th entry $\sigma_{n,m}$. Let $a=(a_1,\ldots,a_N)^T\in\mathbb{R}^N$. Also let $\tilde{1}=(1,\ldots,1)^T\in\mathbb{R}^N$. The

starting point is $E\, a^T X = a^T \mu$ and $\mathrm{Var}(a^T X) = a^T \Sigma a$, which you can get by looking at the matrix formulation (with enough blackboard space). Let $\Delta = \{a \geq 0 : a^T \tilde{1} = 1\}$. The efficient portfolio problem for the rate of return r is

$$\text{Problem } r\text{:} \quad \min_{a \in \Delta} a^T \Sigma a \quad \text{subject to} \quad a^T \mu \geq r.$$

This has a dual problem for variance v,

$$\text{Problem } v\text{:} \quad \max_{a \in \Delta} a^T \mu \quad \text{subject to} \quad a^T \Sigma a \leq v.$$

7.9. **Problems.**

Problem 7.1. *Casella & Berger, 4.5.*

Problem 7.2. *Casella & Berger, 4.7.*

Problem 7.3. *Casella & Berger, 4.13.*

Problem 7.4. *Casella & Berger, 4.26 and 4.27.*

Problem 7.5. *Casella & Berger, 4.41, 4.42, 4.43, and 4.44.*

Problem 7.6. *Casella & Berger, 4.45, 4.46, and 4.47.*

Problem 7.7. *Casella & Berger, 4.58.*

Problem 7.8. *Casella & Berger, 4.62.*

8. Sampling Distributions and Normal Approximations

Problem 8.1. *Casella & Berger, 5.1.*

Problem 8.2. *One of the problems with phone surveys is that they select from people who answer the phone and are willing to answer questions, often personal questions, asked by a stranger. In the context of selection, evaluate the statement, overheard in a Scottish pub, "When a Scotsman moves from Scotland to England, he improves the average IQ in both places."*

Problem 8.3. *The millions of SAT math scores of the population of U.S. collge-bound seniors in* 1978 *(sorry for the old data here) were approximately normally distributed with a mean of* 470 *and a standard deviation of* 120*.*

1. *For a student drawn at random, what is the probability of a score greater than* 500*? Than* 550*? Than* 650*? Than* 810*?*
2. *Averaging a million numbers is easier than it used to be, but I wouldn't want to do it on a calculator. If the mean is estimated from a random sample of* 250*, what is the probability that* $\bar{X}$ *will be no more than* 10 *points off?*
3. *If the mean is estimated from a random sample of* 500*, what is the probability that* $\bar{X}$ *will be no more than* 5 *points off?*
4. *In the last two problems, I asked you to double the sample size and to simultaneously halve the range. Why wasn't the answer the same?*

Problem 8.4. *"One pound" packages are filled by an older machine with weights that vary normally around a mean of* 16.3 *ounces and have a standard deviation of* 0.24 *ounces. An inspector randomly samples* n *packages, weighing them carefully, and fines the company if the sample mean is below* 16 *ounces.*

1. *What is the probability of a fine if* $n = 10$*? If* $n = 20$*? If* $n = 100$*?*
2. *Find the smallest* n *such that the probability of a fine is below* 0.01*.*
3. *If* $n = 50$*, what is the lowest the mean can be and have the probability of a fine at or below* 0.05*? At or below* 0.01*?*

Problem 8.5. *A commuter plane has a rated capacity of* $7,800$ *pounds. The mean passenger weight is* 150 *pounds and the standard deviation is* 30 *pounds.*

1. *If the airline puts* 50 *seats on the plane, what is the probability that the plane's rated capacity will be exceeded on a fully booked flight?*
2. *What is the maximum number of seats the airline can put on the plane to have the probability of exceeding capacity below* 0.001*?*
3. *Suppose that the* 14 *person travelling squad for the basketball team books seats. Using reasonable numbers, recalculate the answer to the previous two questions.*

Problem 8.6. *The old treatment has a survival rate of* 0.56*. In a sample of* 120 *patients, a new treatment had a survival rate of* 0.59*.*

1. *If the new treatment is no better than the old one, how likely is it that the survival rate would have been* 0.59 *or higher?*
2. *Suppose that patients, not knowing the survival rate of the new treatment, were allowed to choose which treatment to receive. How might that change your calculations in the previous problem?*

Problem 8.7. *Casella & Berger, 5.2.*

Problem 8.8. *Casella & Berger, 5.3.*

Problem 8.9. *Casella & Berger, 5.4.*

Problem 8.10. *Any* 2 *of the following problems: Casella & Berger, 5.11, 5.15, 5.17, 5.29, 5.34, 5.44.*

9. Sufficient Statistics as Data Compression

Remember: a statistic, $T(\mathbf{X})$, is a function of the data only, it cannot depend, functionally, on the parameter to be estimated, though we expect its distribution to depend on θ. E.g. $T(\mathbf{X}) \equiv \mathbf{X}$ is a statistic, as is $T(\mathbf{X}) \equiv 7$.

Useful statistics compress data, tell us the important things in the data and suppress the irrelevant details. In this section we look at what are called **sufficient statistics**. We'll also look ahead to the Rao-Blackwell theorem, a truly remarkable result about minimal variance unbiased estimators.

9.1. Sufficient statistics.

When faced with several hundred variables collected on several thousand people over a period of 30 years, one can be excused from thinking that $\mathbf{X}$ is useless because it's too informative, or that 7 is useless because it contains too little information. A good notion of exactly the right amount of information is "everything we need to know about $\mathbf{X}$ in order to make inferences about θ."

Suppose that $\mathbf{X}$ is iid P_θ for some $\theta \in \Theta$, and that the density of $\mathbf{X}$ is $f(\mathbf{x}|\theta)$.

Definition 9.1.1. *A statistic $T(X)$ is **sufficient for** θ if $P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$ does not depend on θ.*

In words, after conditioning on $T(\mathbf{X}) = T(\mathbf{x})$, there is no more probabilistic information about θ in the sample. This is only useful when the dimensionality of the range of T is smaller, typically much smaller, than the number of data points.

Example 9.1.2. *$(X_1, \ldots, X_n) \sim$Binomial(n, p), $p \in (0, 1)$, $T(\mathbf{X}) = (\sum X_i)/n$ is a 1-dimensional sufficient statistic, as is $T'(\mathbf{X}) = e^{(\sum X_i)/n}$. Here is another sufficient statistic, 2-dimensional this time, $T''(\mathbf{X}) = ((\sum X_i)/n, (\sum X_i X_{i+1})/n)$.*

The extra information in $(\sum X_i X_{i+1})/n)$ is not useful **IF** the X_i are iid Bernoulli(p). However, if we are entertaining some doubt about (say) the independence of X_i and X_{i+1}, this is a very informative statistic.

Sometimes we can tell that T is a sufficient statistic from a ratio of conditional probabilities.

Theorem 9.1.3. *If $T(\mathbf{X})$ is sufficient for θ, then for all possible values $\mathbf{x}$, of $\mathbf{X}$, the ratio $p(\mathbf{x}|\theta)/q(T(\mathbf{x})|\theta)$ does not depend on θ.*

Proof: Expand $P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$ using Bayes' Law. ∎

When $X_1, \ldots, X_n$ are iid $N(\mu, 7)$, $\mu \in \mathbb{R}$, then the sample mean is sufficient for μ, and this is true for all values of 7. To see why, rearrange the density of $X_1, \ldots, X_n$ by adding and subtracting $\overline{x}$ and getting a $n(\overline{x} - \mu)$ term, cancel that using the known distribution of $\overline{x}$.

That last was hard work. Sometimes we gain insight into what statistics are sufficient by rearranging the likelihood function so that it factors in the right way.

Theorem 9.1.4 (Halmos-Savage Factorization)**.** *$T(\mathbf{X})$ is sufficient for θ iff there exists a function $(t,\theta) \mapsto g(t|\theta)$ and a function $x \mapsto h(x)$ such that*

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$$

for all t in the range of T, all $\theta \in \Theta$, and all x in the range of $\mathbf{X}$.

Proof: If $T(\mathbf{X})$ is sufficient, then

$$f(\mathbf{x}|\theta) = P_\theta(\mathbf{X}=\mathbf{x}) = P_\theta(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x}))P_\theta(T(\mathbf{X})=T(\mathbf{x})) = h(\mathbf{x})g(T(\mathbf{x})|\theta).$$

Now suppose that $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$. Let $q(t|\theta)$ be the pmf for T and let $A_x = \{\mathbf{y} : T(\mathbf{y}) = T(\mathbf{x})\}$.

$$\begin{aligned}\frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\sum_{\mathbf{y}\in A_x} g(T(\mathbf{y})|\theta)h(\mathbf{y})} \\ &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{g(T(\mathbf{x})|\theta)\sum_{\mathbf{y}\in A_x} h(\mathbf{y})} \\ &= \frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_x} h(\mathbf{y})},\end{aligned}$$

which does not depend on θ. ∎

Two observations:

1. If $T(\mathbf{X})$ is a sufficient statistic for θ, $\widehat{\theta}_{MLE}$ is a function of the data **only** through T.
2. The Halmos-Savage theorem gives us, essentially for free, the following observation:
   If $f(\mathbf{x}|\theta) = h(\mathbf{x})c(\theta)\exp\left(\sum_{i=1}^k w_i(\theta)t_i(\mathbf{x})\right)$, then

$$T(\mathbf{X}) = \left(\sum_{j=1}^n t_1(X_j), \ldots, \sum_{j=1}^n t_k(X_j)\right)$$

   is a sufficient statistic for θ, $\theta \in \mathbb{R}^d$, $d \le k$.

9.2. **Rao-Blackwell.** Some time ago we looked at properties of estimators. In particular, we used the Cauchy-Schwarz inequality and some fancy side-stepping to show that we had found a minimum variance unbiased estimator in the class of all estimators. Here's another way to get at such wonderfulness. I said it before, but I'll say it again, an estimator must be a function of the data, and not a function of θ.

Recall that $E[E(X|Y)] = E\,X$ and $\operatorname{Var} X = \operatorname{Var}[E(X|Y)] + E[\operatorname{Var}(X|Y)]$.

Theorem 9.2.1 (Rao-Blackwell)**.** *Let $\widehat{\theta}$ be any unbiased estimator of θ and let T be a sufficient statistic for θ. Define $\varphi(T) = E(\widehat{\theta}|T)$. Then for all θ, $E_\theta\varphi(T) \equiv \theta$ and $Var_\theta\varphi(T) \le Var_\theta\widehat{\theta}$.*

Since the result holds for **any** unbiased estimator, what we know is that $\varphi(T)$ minimizes variance amongst all unbiased estimators.

Proof: For unbiasedness,

$$\theta \equiv E_\theta \widehat{\theta} = E_\theta[E(\widehat{\theta}|T)] = E_\theta \varphi(T).$$

For minimal variance,

$$\begin{aligned} \mathrm{Var}_\theta \widehat{\theta} &= \mathrm{Var}_\theta[E(\widehat{\theta}|T)] + E_\theta[\mathrm{Var}(\widehat{\theta}|T)] \\ &= \mathrm{Var}_\theta(\varphi(T)) + E_\theta[\mathrm{Var}(\widehat{\theta}|T)] \\ &\geq \mathrm{Var}_\theta(\varphi(T)). \end{aligned}$$

Finally, since T is a sufficient statistic for θ, after conditioning on T, the result cannot depend on θ, which shows that $\varphi(T)$ is really an estimator. ∎

Example 9.2.2. *$X_1, \ldots, X_n$ iid $U(0,\theta)$, show that $T := \max(X_1, \ldots, X_n)$ is sufficient, find the unbiased function of T.*

At the beginning of the prob/stats part of this course, we used the information inequality to show that $\widehat{p}$ being the sample proportion achieves the minimal variance amongst all unbiased estimators. We can do that a bit more easily now . . . .

9.3. **Problems.**

Problem 9.1. *Casella & Berger, 6.1.*

Problem 9.2. *Casella & Berger, 6.3.*

Problem 9.3. *Casella & Berger, 6.4.*

Problem 9.4. *Casella & Berger, 6.9.*

Problem 9.5. *Any three other problems from Casella & Berger, Ch. 6.1.*

Problem 9.6. *Casella & Berger, 7.19, 7.20, and 7.21.*

10. Finding and Evaluating Estimators

The basic idea is to guess the true θ, that is, to form an **estimator** $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$. Any function, $W = W(X_1, \dots, X_n)$ is called a **statistic**. Statistics that are supposed to guess at some true θ, that is, estimators, are the ones we study most often.

Why estimate some θ in a set Θ rather than trying to work directly with getting the true distribution? There are (at least two good reasons, parsimony, and θ may have some intrinsic meaning that we care about.

1. Parsimony — this becomes much more important in higher dimensions. Given $X_1, \dots, X_n \in \mathbb{R}^k$, define the empirical cdf by $F_n(x) = \frac{1}{n}\sum_{i \le n} 1_{(-\infty, x]}(X_i)$. Think about how many points you need to form an ϵ-net for $[0, 100]^k$, e.g. $\epsilon = 0.01$, $k = 18$, the answer is, roughly,

   $$1,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,$$

   which is a pretty big number. As data grows, the empirical cdf will get closer and closer to the true cdf. But it will do so more and more slowly when the number of dimensions is large.

   Compare this to $F_{\hat{\theta}_n}(x) = \int \cdots \int_{-\infty}^{x} f(y|\hat{\theta}_n)dy$. The first is as close to the data as possible, the second is derived from getting "as close to" θ as possible.
2. The parameter θ may have some intrinsic meaning, like the wage elasticity of labor supply or the gravitational constant.

10.1. **The basic Gaussian example.** $X_1, \dots, X_n$ are iid with distribution belonging $\{f(x|(\mu, \sigma^2)) : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{++}\}$ where

$$f(x|(\mu, \sigma^2)) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}.$$

The basic statistics we form are

1. $\overline{X}_n = \frac{1}{n}\sum_{i \le n} X_i$, an estimator of μ, and
2. $S^2 = \frac{1}{n-1}\sum_{i \le n}(X_i - \overline{X}_n)^2$, an estimator of σ^2.

Note that $\overline{X}$ and S^2 are unbiased, and that the two statistics are sufficent for the normal distribution. That makes the look pretty good. They're also fairly easy to use, thanks to the following.

Claim: $\overline{X}_n \sim N(\mu, \sigma^2/n)$, $nS^2/\sigma^2 \sim \chi^2(n-1)$, and $\overline{X}_n \perp\!\!\!\perp S^2$.

Using this claim is one of the basic set of competencies to be learned in an introductory prob/stat class.

10.1.1. *Intervals around our guess of μ, σ known.* We've used $\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim N(0, 1)$ a number of times to get at intervals in which we believe it most likely that μ lies.

10.1.2. *Intervals around our guess of σ.* If $Z_1, \dots, Z_r$ are iid $N(0,1)$, then the random variable $X = \sum_{i \leq r} Z_i^2$ has a $\chi^2(r)$ distribution. Note that $E\,X = r$, and X is the sum of independent random variables. Therefore, for moderately large r, $\frac{1}{\sqrt{r}}(X - r)$ is approximately normal. The approximation is off for a couple of reasons, the easiest of which is that $X > 0$ always (being the sum of squares of independent, non-degenerate rv's), while the approximation allows $X < 0$. This is why one most often uses the explicit tabulations of the χ^2 random variables.

$nS^2/\sigma^2 = \frac{n}{n-1} \sum_{i \leq n} (X_i - \overline{X}_n)^2/\sigma^2 \sim \chi^2(n-1)/\sigma^2$. Let $X \sim \chi^2(n-1)$, from a table, find $r' < r''$ so that $P(X \leq r') = 0.01$ and $P(X \leq r'') = 0.99$ so that $P(r' \leq X \leq r'') = 0.98$. Then

$$P\left(\frac{n}{n-1} \sum_{i \leq n} (X_i - \overline{X}_n)^2/\sigma^2 \leq r'\right) = P\left(S^2 \leq \frac{\sigma^2 \cdot r'}{n}\right) = 0.01,$$

$$P\left(\frac{n}{n-1} \sum_{i \leq n} (X_i - \overline{X}_n)^2/\sigma^2 \leq r''\right) = P\left(S^2 \leq \frac{\sigma^2 \cdot r''}{n}\right) = 0.99,$$

$$P\left(r' \leq \frac{n}{n-1} \sum_{i \leq n} (X_i - \overline{X}_n)^2/\sigma^2 \leq r''\right) = P\left(\frac{\sigma^2 \cdot r'}{n} \leq S^2 \leq \frac{\sigma^2 \cdot r''}{n}\right) = 0.98.$$

This expresses the intervals as proportions of the true σ^2. If we take square roots all over the place, this expresses the intervals as proportions of the true σ.

Do some examples of using the tabulated χ^2.

10.1.3. *Intervals around our guess of μ, σ <u>not</u> known.* If $Z \sim N(0,1)$, $V \sim \chi^2(r)$ and $Z \perp\!\!\!\perp V$, then the name of the random variable

$$T = \frac{Z}{\sqrt{V/r}}$$

is *Student's t with r degrees of freedom*, name after the famous (not!) Guiness brewery employee, W. S. Gosset. Going back to the claim,

Claim: $\overline{X}_n \sim N(\mu, \sigma^2/n)$, $nS^2/\sigma^2 \sim \chi^2(n-1)$, and $\overline{X}_n \perp\!\!\!\perp S^2$,

we just need to scale $\overline{X}_n$ and $nS^2/\sigma^2 \sim \chi^2(n-1)$, and form their ratio to get a t distribution with $n-1$ degrees of freedom. Look at the algebra when we do that, the unknown σ^2 disappears. This is what we use to form intervals when we replace the unknown σ by a good guess, $\hat{\sigma} = \sqrt{S^2}$.

10.2. **Some examples of finding estimators.** We're going to start with a number of examples where we want to find a parameter of interest, then we'll turn to two different methods of finding estimators, the method of moments and maximum likelihood.

10.2.1. *Examples.* Roughly, I divide examples into those in which we observe everything relevant to us in the sample and those we do not. This latter category is where research interest in economics often centers.

1. A new medical treatment keeps people alive for two years with probability p and we'd like to know p. Here $X_1, \ldots, X_n$ are iid Bernoulli(p).
2. I observe that this year's first year class seems taller than the previous years. Rather than measure the heights of all 18,000 first and second year students, I sample 100 of each, wanting to know the heights of the people in the two years. Here I might model $X_{n,y} \sim N(\mu_y, \sigma^2)$, $y = 1, 2$, and I want to estimate $\mu_1 - \mu_2$.
3. In each of the last n years, k crimes of a given type were committed, and with probability p, the crime was reported. Given only the reported number, $X_1, \ldots, X_n$, I want to estimate both k and p. Here the X_n are iid Binomial(k, p) and I know neither k nor p.
4. I want to know the average life of a batch of chips. I test n of them for 6 months, $n_f \leq n$ of them have failed at time $\tau_n \leq 6$, the rest have not, $\tau_n > 6$. Here it is reasonable to assume that the τ_n are iid exponential(λ).
5. We observe the spending of $\$X_n$ on person n in a job training program. We observe whether or not they have a job 1 year later. We'd like to know $\partial P(\text{job in 1 year})/\partial \$$.
6. We observe years of schooling Y_n for person n and their wage rate at 40 years of age, W_n. We'd like to know $\partial W/\partial Y$. There are hidden variables at work, just as there are in the classic fertilizer studies.
7. I am building a shark cage and want to save money by making it only as strong as it needs to be to survive shark attack by the biggest shark. The measurements of the sharks that I have killed so far are X_n. I want to estimate the maximum shark size.
8. The monsoons will come at some random time T in the next month, $T \in [0, 1]$. A farmer must pre-commit to a planting time, a. As a function of a and t, the realized value of T, the harvest will be

$$h(a,t) = \begin{cases} K - r|a-t| & \text{if } a \leq t \\ K - s|a-t| & \text{if } a > t \end{cases}$$

where $K > r, s > 0$. The random arrival time of the monsoon has cumulative distribution function $F(\cdot)$, and $f(x) = F'(x)$ is its strictly positive probability density function. We've observed $T_1, T_2, \ldots, T_n$ monsoon arrival times in the past, and want to estimate the optimal time.

10.2.2. *Method of moments.* Assume the basic statistical model with a k-dimensional parameter set, that is, suppose that $\theta \in \mathbb{R}^k$. Solve for k equations in k unknowns $M(\theta) = M(X_1, \ldots, X_n)$, $M(\cdot)$ being "moments implied by the argument," that is, for the θ giving

the k first moments of the data. Why stop at k moments? Why use integer moments? No good reason. However, people become uneasy when you tell them that that are zillions and zillions of things they could do.

E.g. $N(\mu, \sigma^2)$, Binomial(k, p), $U(0, \theta)$ i.e. $f(x|\theta) = \frac{1}{\theta}$, $0 < x < \theta$ (one version of the shark problem), exponential(β).

10.2.3. *Maximum likelihood.* E.g. $N(\mu, \sigma^2)$, $U(0, \theta)$ (one version of the shark problem), $f(x|\theta) = \theta x^{-2}$, $0 < \theta \leq x < \infty$, exponential$(\beta)$, two-tailed exponential location family.

10.3. **Problems.**

Problem 10.1. *Casella & Berger, 7.1.*

Problem 10.2. *Casella & Berger, 7.6.*

Problem 10.3. *Casella & Berger, 7.10.*

Problem 10.4. *Casella & Berger, 7.13.*

Problem 10.5. *Casella & Berger, 7.14.*

11. Evaluating different estimators

Readings: Casella & Berger, Chapter 7.3, and these notes.

One of the many things we've learned in the previous sections is that, even for a given statistical model $X_1, \ldots, X_n$ iid P_θ, $\theta \in \Theta$, there are lots of estimators of θ. We need a way to distinguish between them. The main one is mean squared error.

11.1. **Mean Squared Error (MSE).** I'd like to give three **really stupid** estimators, $\hat{\theta}_{rs}$, $\hat{\theta}'_{rs}$, and $\hat{\theta}''_{rs}$, and one pretty good one, $\hat{\theta}_{MLE}$, to indicate the kinds of things we'd like to be able to rule out. We'll then look at how MSE works for these estimators.

Three really stupid estimators, and one pretty good one:

1. $X_1, \ldots, X_n$ iid P_θ, $P_\theta = N(\theta, 1)$, $\theta \in \mathbb{R}$, $\hat{\theta}_{rs} = X_1$.
2. $X_1, \ldots, X_n$ iid P_θ, $P_\theta = N(\theta, 1)$, $\theta \in \mathbb{R}$, $\hat{\theta}'_{rs} = 7$.
3. $X_1, \ldots, X_n$ iid P_θ, $P_\theta = N(\theta, 1)$, $\theta \in \mathbb{R}$, $\hat{\theta}''_{rs} = 2 \cdot \frac{1}{n} \sum_{i=1}^n X_i$.
4. $X_1, \ldots, X_n$ iid P_θ, $P_\theta = N(\theta, 1)$, $\theta \in \mathbb{R}$, $\hat{\theta}_{MLE} = \frac{1}{n} \sum_{i=1}^n X_i$.

A general criterion for picking between estimators so as to make some, hopefully sensible, kind of tradeoff between bias and variance is called Mean Squared Error (MSE). It is

$$\mathrm{MSE}_\theta(\hat{\theta}) := E_\theta(\hat{\theta} - \theta)^2.$$

The basic result about MSE's is

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{Bias}^2_\theta(\hat{\theta}) + \mathrm{Var}_\theta(\hat{\theta}).$$

Do this calculation.

1. $\mathrm{MSE}_\theta(\hat{\theta}_{rs}) = \mathrm{Var}_\theta(\hat{\theta}_{rs}) = 1$ because $E_\theta X_1 = \theta$ so the estimator is unbiased. However, since the estimator throws away the data points $X_2, \ldots, X_n$, it is not sensible, $\mathrm{MSE}_\theta \equiv 1$ does not go to 0 as $n \uparrow \infty$. This means that as we get more and more data, the variance of this estimator will not converge to 0 and the estimator will not converge to the true θ.
2. $\mathrm{MSE}_\theta(\hat{\theta}'_{rs}) = (7 - \theta)^2 + 0$ since the variance of $\hat{\theta}'_{rs}$ is equal to 0. This converges to 0 very very quickly (instantly) if the true θ is 7, but not at all if the true θ is someplace else. Since the variance is always equal to 0, we learn that comparing estimators by comparing only their variance is too narrow.
3. $\mathrm{MSE}_\theta(\hat{\theta}''_{rs}) = \theta^2 + \frac{4}{n}$. Here the variance of the estimator goes to 0, but the estimator is biased unless the true θ is equal to 0. Even if the true θ is equal to 0, the MSE is four times as large as it need be, as can be seen in the next estimator.
4. $\mathrm{MSE}_\theta(\hat{\theta}_{MLE}) = 0^2 + \frac{1}{n}$. Here the MSE does not depend on θ and goes to 0 as $n \uparrow \infty$.

Draw a graph of the MSE's of the four estimators as a function of θ, see what you see, in formal terms, $\leq$ is only a partial order on the set of functions.

11.2. **Desirable properties for estimators.** Some of the properties we might like, but will not insist on, for estimators include:

1. Unbiased estimators. If for all $\theta \in \Theta$,
$$E_\theta \hat{\theta}(X_1, \dots, X_n) = \theta,$$
then you have an **unbiased** estimator. This means that, on average, your estimator is right. $\hat{\theta}_{rs}$ is an example that shows that this is not enough.
2. Best unbiased estimators (BUE's). If you have many unbiased estimators of θ, maybe it makes sense to pick the one with the lowest variance, which, in this case is the same as picking the one with the lowest MSE. Such an estimator is called a BUE. The **relative efficiency** of two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ is given by
$$\frac{\text{Var}(\hat{\theta}_1)}{\text{Var}(\hat{\theta}_2)}.$$
Because we've asked for unbiased, the problem of division by 0 (as in $\hat{\theta}'_{rs}$) does not arise.
3. Best linear unbiased estimators (BLUE's) Another thing you might impose on the estimators is that they be linear functions of the data. This makes the most sense when you're after things like the mean of the data rather than some non-linear function of the data like the variance.

Example: Find the BLUE for the mean of iid mean μ rv's $\{X_i\}_{i=1}^n$.

Example: We're given non-zero numbers $x_1, \dots, x_n$ and the rv's $Y_1, \dots Y_n$ are independent, having the distribution described by
$$Y_i = \beta x_i + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2),$$
β and σ^2 unknown. Notice that the ε_i are iid but the Y_i are not (unless the x_i are all equal to each other). Another way to write this is that the independent Y_i satisfy $Y_i \sim N(\beta x_i, \sigma^2)$. The essential intuition is when $|x_i|$ is large, the i'th observation tells one more about the value of β. Here are three unbiased estimators of β, compare their variances: $\hat{\beta}_{MLE}$ (find it, it's the obvious one); $\hat{\beta}' = (\sum_i Y_i)/(\sum_i X_i)$; and $\hat{\beta}'' = \frac{1}{n}\sum_i (Y_i/X_i)$.

11.3. **The Cramér-Rao lower bound.** Re-examine this. It is a lower bound on the possible variance of **any** unbiased estimator, linear, non-linear. This means that if you've found an unbiased estimator with this variance, you've done the best you possibly can. The only drawback is that, quite often, you cannot do this well, quite often, no estimator can achieve the bound.

11.4. **Problems.**

Problem 11.1. *Casella & Berger, 7.37.*

Problem 11.2. *Casella & Berger, 7.38.*

Problem 11.3. *Casella & Berger, 7.41.*

Problem 11.4. *Casella & Berger, 7.49.*

Problem 11.5. *Suppose that $X_1, \ldots, X_n$ are iid $f(x|\theta)$, $\theta \in \Theta \subset \mathbb{R}_{++}$ and that $\hat{\theta}(X_1, \ldots, X_n)$ is an unbiased estimator of θ, and suppose that neither the X_i not $\hat{\theta}$ are degenerate random variables. By considering the estimators $s \cdot \hat{\theta}$, $0 < s < 1$, show that $\hat{\theta}$ does* **not** *minimize MSE. One way to proceed is to define*

$$f(s) = E(s\hat{\theta} - \theta)^2 = s^2 var(\hat{\theta}) + \theta^2(s-1)^2.$$

This is a quadratic in s that opens upwards. Find its minimum and show that it happens for $s < 1$. [The "s" is a "shrinkage factor," this problem tells us that unbiased estimators do not minimize MSE.]

12. Hypothesis Testing

> Say not "I have found the truth," but rather, "I have found a truth." (Kahlil Gibran, *The Prophet*)

Readings: Casella & Berger, Chapter 8, and these notes.

12.1. **Overview.** We're still working with the basic statistical model, the data $\mathbf{X} = (X_1, \ldots, X_n)$ are iid $f(x|\theta)$, and whenever it's convenient, I'll use $f(\mathbf{x}|\theta)$ or $L(\mathbf{x}|\theta)$ for the likelihood function evaluated at $\mathbf{X} = \mathbf{x}$.

The useful estimators $\hat{\theta}(\mathbf{X})$ are, typically, non-degenerate random variables. This means that they have a real spread, e.g. strictly positive variance. This in turn depends on the true θ. The essential intuitive idea behind hypothesis testing is that we can use this spread to tell how likely or unlikely it is that we would have seen the data that we did see if the true θ belonged to some pre-specified set of interest, call it Θ_0. In particular, we could form a rule "we'll say that the true θ does not belong to Θ_0, that is, we'll **reject** Θ_0, if the data belong to the set $\mathbb{X}_r$ which is unlikely to show up unless the true θ is outside of Θ_0.

The essential ingredients are then

1. The basic statistical model, $\mathbf{X} \sim f(\mathbf{x}|\theta)$, $\theta \in \Theta$,
2. a null hypothesis, $H_0 : \theta \in \Theta_0$, $\Theta_0 \subset \Theta$,
3. the alternative, $H_1 : \theta \notin \Theta_0$,
4. a decision rule, reject H_0 if $\mathbf{x} \in \mathbb{X}_r$ and accept H_0 if $\mathbf{x} \notin \mathbb{X}_r$.

We can then examine the probabilistic properties of the decision rule using the **power function**, $\beta(\theta) = P(\mathbb{X}_r|\theta)$.

12.2. **The perfect power function and types of errors.** The perfect power function is $\beta(\theta) = 1_{\Theta_0^c}(\theta)$, that is, reject if and only if the null hypothesis is false. (Sing a bar of "To dream the impossible dream.") However, the idea behind the basic statistical model is that we do not observe θ directly, rather we observe the data $\mathbf{X}$, and the data contains probabilistic information about θ. In statistics, we don't expect to see perfect power functions, they correspond to having positive proof or disproof of a null hypothesis.[3]

Here's a (contrived) example where the data does entirely distinguish between the null and the alternative.

Example: $\Theta = \{\theta_1, \theta_2\}$, $f(x|\theta_1) = 1_{(0,1)}(x)$, $f(x|\theta_2) = 1_{(2,3)}(x)$. The rejection region $\mathbb{X}_r$ is the set of $\mathbf{x}$ where each x_i satisfies $2 < x_i < 3$. This is, or should be, obvious and straightforward.

[3]This is why people are often so dissatisfied with statisticians, remember Mark Twain's saying about them, "There are liars, there are damned liars, and there are statisticians."

Most of the time, things are more difficult, and our data will leave open the possibility that we are making some kind of mistake. There are two types of errors that you can make, unimaginatively called **Type I** and **Type II** errors:

1. you can reject a null hypothesis even though it is true, this kind of false rejection of a true null hypothesis is called a **Type I** error.
2. you can accept the null hypothesis even though it is false, this kind of false acceptance of a null hypothesis is called a **Type II** error.

The following table may help:

Decision\ "Truth"	$H_0 : \theta \in \Theta_0$	$H_1 : \theta \notin \Theta_0$
Accept H_0	Bingo!	Type II
Reject H_0	Type I	Bingo!

We're going to look at errors we make in the context of the following two examples.

Example: $\Theta = [0,1]$, $X_1, \ldots, X_n$ are iid Bernoulli(θ), $H_0 : \theta \leq \frac{1}{2}$, $H_1 : \theta > \frac{1}{2}$. Intuitively, the rejection region $\mathbb{X}_r$ should be the set of $\mathbf{x}$ such that the sample average, $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n x_i$ belongs to the interval $(\frac{1}{2} + c, 1]$ for some number $0 < c < \frac{1}{2}$.

This corresponds to the story about comparing a new medical procedure with an old one that had a 50% success rate. The null hypothesis is that the new treatment has no better than the old one, we reject this null hypothesis if the evidence strongly favors the new one. In particular, we reject the null if the sample average success rate is above $\frac{1}{2}+c$. The higher is c, the lower is the probability of a false rejection but the higher is the probability of a false acceptance.

Example: $\Theta = [0,1]$, $X_1, \ldots, X_n$ are iid Bernoulli(θ), $H_0 : \theta \in \{\frac{1}{2}\}$, $H_1 : \theta \neq \frac{1}{2}$. Intuitively, the rejection region $\mathbb{X}_r$ should be the set of $\mathbf{x}$ such that the sample average, $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n x_i$ belongs to the pair of intervals $[0, \frac{1}{2} - a) \cup (\frac{1}{2} + b, 1]$ for some pair of numbers $0 < a, b < \frac{1}{2}$. Perhaps we even expect that $a = b$.

Before we do any calculations, we should think through

12.3. **Some generalities about the probabilities of the different types of errors.** The power function will help us look at the probabilities of the different kinds of errors. It should be intuitively clear that, if we're using sensible procedures, lowering the probability of false rejection entails raising the overall probability of acceptance, including the probability of making a false acceptance.

The number $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$ gives the highest probability of a Type I error. This is called the **size** of the test $\mathbb{X}_r$. $1 - \alpha$ is the confidence level of the test. Confidence is good, we'd like our confidence to be close to 1. We'd also like a powerful test. Follow through the logic of the next two:

1. If we increase $\mathbb{X}_r$, we lower α, raising both our confidence and decreasing the power of our test.
2. If we decrease $\mathbb{X}_r$, we raise α, lowering our confidence but increasing the power of our test.

Now, instead of increasing or decreasing $\mathbb{X}_r$, we could add some points and take away other points. If we can do that so as to simultaneously increase both confidence and power (that is, simultaneously reduce both α and β), then our original test was stupid. In the perfect power function example above, $\alpha = \beta = 0$, we were completely confident with our most powerful possible test. In general there is a tradeoff between confidence and power.

In tabular form

Decision\"Truth"	$H_0 : \theta \in \Theta_0$	$H_1 : \theta \notin \Theta_0$
Accept H_0	$(1-\alpha) =$ confidence	Type II (β)
Reject H_0	Type I (α)	$(1-\beta) =$ power of test

Go through the Wonnacott and Wonnacott Gaussian one-sided hypothesis testing pictures identifying confidence and power as areas under curves.

12.4. **The Likelihood Ratio Tests.** The number

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta\in\Theta_0} L(\mathbf{x}|\theta)}{\sup_{\theta\in\Theta} L(\mathbf{x}|\theta)}$$

must be in the interval $[0,1]$. The class of likelihood ratio tests (LRTs) are

$$\mathbb{X}_r(c) = \{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}.$$

In the perfect power function example, for any $0 < c < 1$, the LRT is the same, and is the one we gave above.

In general, the smaller is c, the smaller is $\mathbb{X}_r(c)$, the larger is α and the smaller is β.

Example: $\Theta = [0,1]$, $X_1, \ldots, X_n$ are iid Bernoulli(θ), $H_0 : \theta \in \{\frac{1}{2}\}$, $H_1 : \theta \neq \frac{1}{2}$. Find the LRTs as a function of c and n.

Detour through properties of the function $f(x) = x\log x$: If $x > 0$, then $f(x) = x\log x$ is well defined. By calculation, $f^{(1)}(x) = \log x + 1$ which has a critical point at $x^* = \exp(-1)$. Further, $f^{(2)}(x) = x^{-1} > 0$. So, on $(0,\infty)$, $f(x)$ is strictly convex achieving its global minimum at $\exp(-1)$. To calculate behavior around $x = 0$, consider the following two substitutions:

$$\lim_{x\to 0} x\log x = \lim_{y\to-\infty} \exp(y)\log(\exp(y)) = \lim_{z\to+\infty} \frac{-z}{\exp(z)} = 0 - .$$

Hence, $f(0)$ can be defined equal to 0 by continuity.

Alright, let's calculate the LRTs for the example. $L(\mathbf{x}|\theta) = \Pi_{i=1}^n \theta^{x_i}(1-\theta)^{(1-x_i)}$. Therefore,

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta\in\Theta_0} L(\mathbf{x}|\theta)}{\sup_{\theta\in\Theta} L(\mathbf{x}|\theta)} = \text{etc.}$$

After doing some algebra, the rejection region is the set of $\mathbf{x}$ such that the sample average, $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n x_i$ satisfies

$$\overline{\mathbf{x}}\log(\overline{\mathbf{x}}) + (1-\overline{\mathbf{x}})\log(1-\overline{\mathbf{x}}) \geq \log\tfrac{1}{2} - \frac{\log c}{n}.$$

Note that the lhs of the inequality is symmetric about $\frac{1}{2}$, that $\overline{\mathbf{x}} = \frac{1}{2}$ never satisfies the inequality for $0 < c < 1$, and that the rejection region is of the form pair of intervals $[0, \frac{1}{2} - a) \cup (\frac{1}{2} + a, 1]$ for some $0 < a < \frac{1}{2}$. Find $\beta(\theta)$, α and β.

Example: $\Theta = [0,1]$, $X_1, \ldots, X_n$ are iid Bernoulli(θ), $H_0 : \theta \leq \frac{1}{2}$, $H_1 : \theta > \frac{1}{2}$. Find the LRTs as a function of c and n.

Example: $X_1, \ldots, X_n$ are iid $N(\theta, 1)$, $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$. Find the LRTs as a function of c and n.

Example: We see iid $X_1, \ldots, X_n$ and iid $Y_1, \ldots, Y_m$, both from Gaussian populations with the same variance. The null hypothesis is that the means of the two populations are the same. What do the tests look like?

Example: We see iid $X_1, \ldots, X_n$ and iid $Y_1, \ldots, Y_m$, both from populations with the same variance. The null hypothesis is that the means of the two populations are the same. Now, provided n and m are large enough for the CLT, what do the tests look like?

12.5. **Confidence intervals, p-values, and hypothesis testing.** Let $X_1, \ldots, X_n$ be iid with unknown mean μ and unknown variance σ^2. The $(1-\alpha)$ confidence interval around μ was calculated as $[\overline{X} - t_{\alpha/2,n}S, \overline{X} - t_{\alpha/2,n}S]$ where $\overline{X}$ is our estimator of μ and S is our estimator of σ.

The null hypothesis is $\mu = \mu_0$, that is, $\Theta_0 = \{\mu_0\} \subset \mathbb{R}$. We accept H_0 at the confidence level $(1-\alpha)$ if

$$[\overline{X} - t_{\alpha/2,n}S, \overline{X} - t_{\alpha/2,n}S] \cap \Theta_0 \neq \emptyset,$$

that is, if

$$\mu_0 \in [\overline{X} - t_{\alpha/2,n}S, \overline{X} - t_{\alpha/2,n}S],$$

otherwise we reject H_0 at the α level.

Typically, α is one of the following list of numbers: 0.1, 0.05, 0.01, or 0.001. Suppose that we've rejected at one of these α levels, say 0.05. What about the other levels? Well, we know ....

The p-value gets around the problem of how to report rejection/acceptance. Specifically, we find the smallest number r such that

$$\mu_0 \in [\overline{X} - t_{r/2,n}S, \overline{X} - t_{r/2,n}S],$$

and this is the p-value.

12.6. **Problems.**

Problem 12.1. *Casella & Berger, 8.1.*

Problem 12.2. *Casella & Berger, 8.2.*

Problem 12.3. *Casella & Berger, 8.3.*

Problem 12.4. *A poll asked a number of Americans whether or not atomic power plants are safe enough. Of the* 420 *aged between* 18 *and* 30, 24% *answered "Yes," of the* 510 *aged between* 30 *and* 50, 34% *answered "Yes." H_0 is the hypothesis that age makes no difference.*

1. *Calculate the p-value for H_0.*
2. *At level $\alpha = 0.05$, can H_0 be rejected?*

Problem 12.5. *Yields of plants of a new high-yield variety of wheat are approximately normally distributed with a standard deviation of* 15 *grams. A researcher plans to grow a sample of n plants.*

1. *What is the minimum sample size required for the probability to be* 0.99 *that the sample mean will fall within* 1 *gram of the population mean.*
2. *Why might a researcher want to base calculations on n different acres, planted at some distance from each other, rather than n different plants?*

Problem 12.6. *Casella & Berger, 8.12.*

Problem 12.7. *Casella & Berger, 8.13.*

Problem 12.8. *Casella & Berger, 8.14.*

Problem 12.9. *Casella & Berger, 8.15.*

Problem 12.10. *Suppose that $X_1, \ldots, X_n$ are i.i.d. with the uniform distribution on $(0, \theta)$ for some unknown $\theta \in \Theta = (0, +\infty)$.*

1. *Give the likelihood function $L(X_1, \ldots, X_n : \theta)$.*
2. *Give the maximum likelihood estimator (MLE) of θ.*
3. *Find the 95% confidence interval for the MLE.*
4. *Calculate the bias of the MLE.*
5. *Describe the $\alpha = 0.01$ one-sided test for the null hypothesis $\theta \geq 19$.*
6. *Calculate the power of the test as a function of $\theta < 19$.*