Printed: 27 February 2022

Chapter 2

It's just Calculus and convexity

Calculus::Calculus

SECTION 2.1 *explains why you can safely just skim through this chapter on a first pass.*

SECTION 2.2 *Describes a simple Taylor expansion trick for bounding functions of the form* $2(f(x) - f(0) - xf'(0))/x^2$. *I have found the trick a very good way of avoiding masses of Calculus. Two important special cases are described.*

SECTION 2.3 *explains why the logarithm of a moment generating function (MGF) is always convex.*

SECTION 2.4 *describes briefly a procedure known as the Fenchel-Legendre transformation, explaining how it relates to a minimization method (Chapter 3) for deriving probability tail bounds. The magical* ψ_{Benn} *makes its first appearance.*

SECTION 2.5 *waxes historical about the Stirling formula and binomial coefficients. It describes a few tricks that pop up often in the literature on high-dimensional statistics.*

SECTION 2.6 *is an advertisement for the convex function* ψ_{Benn}. *It has been a hobby of mine to find hidden appeals to the helpful properties of that function buried inside complex arguments.*

SECTION 2.7 *explains why many important quantities are written as integrals.*

2.1 Introduction

Calculus::S:intro

You could safely skim quickly through this chapter, just to get an idea of how much fancy modern theory depends on only a little bit of Calculus with magical inequalities coming via convexity arguments, sometimes in disguise. I'll tell you in later chapters when something from here is needed, at which point you might want to read some sections from this Chapter more carefully. I prefer having the elementary—but very useful—arguments collected together in one place rather than having them buried inside complicated proofs scattered all over the place.

2.2 Taylor's theorem

Calculus::S:Taylor

Where would Mathematics be without approximation of functions by polynomials?

For my purposes, the integral form of Taylor's theorem is the most useful. Suppose g is a real-valued function defined at least on an interval J of the real line that contains both 0 and a point x. If g is twice continuously differentiable then

$$g(x) - g(0) = \int_0^1 \frac{\partial g(xt)}{\partial t}\, dt = x \int_0^1 g'(xt)\, dt \qquad \text{for } x \in J$$

and

$$\begin{aligned} R_1(x,g) &= g(x) - g(0) - xg'(0) \\ &= x \int_0^1 g'(xt) - g'(0)\, dt \\ &= x^2 \int_0^1 \int_0^1 \{0 \le s \le t \le 1\} g''(xs)\, ds\, dt \\ &= \tfrac{1}{2} x^2 \Delta(x,g) \qquad \text{where } \Delta(x,g) := \int_0^1 2(1-s) g''(xs)\, ds \end{aligned}$$

\E@ R1.rep <1>

And so on, if higher-order derivatives exist. See Problem [1] for the analog when g has an absolutely continuous kth derivative.

Remark. The assumptions on g could be relaxed slightly by using some fancy facts about absolute continuity.

The behaviors of $R_1(x,g)$ and $\Delta(x,g)$ are controlled by g'' . For example, if g'' is non-negative then so are $\Delta(\cdot)$ and $R_1(\cdot)$, and if g'' is continuous then so are $\Delta(\cdot)$ and $R_1(\cdot)$. More interestingly, if g'' is a nondecreasing function then so is $\Delta(\cdot)$, and if g'' is a convex function then so is $\Delta(\cdot)$.

We can interpret the factor $2(1-s)$ in the definition of Δ as the density (with respect to Lebesgue measure) of a probability measure on the interval $(0,1)$, with expected value

$$\int_0^1 s2(1-s)\,ds = 1/3.$$

If g'' is convex then Jensen's inequality gives

\E@ integral.Jensen <2>
$$\Delta(x,g) \geq g''\left(x\int_0^1 2s(1-s)\right) = g''(x/3).$$

This inequality accounts for a few factors of $1/3$ that pop up in well known inequalities.

My main interest in these Calculus tricks originally came from studying two special examples, which will be important throughout this book. (Hence the funny fonts.) The two functions are

\E@ fbb <3>
$$\mathbb{f}(x) = e^x - 1 - x \qquad \text{for } x \in \mathbb{R}$$

\E@ hbb <4>
$$\mathbb{h}(x) = (1+x)\log(1+x) - x \qquad \text{for } x \geq -1.$$

For $\mathbb{h}$ the convention $0\log 0 = 0$ gives $\mathbb{h}(-1) = 1$. Sometimes the domain of $\mathbb{h}$ is extended to the whole real line by defining $\mathbb{h}(x) = +\infty$ when $x < -1$, for reasons that will be explained in the next Section.

The first and second derivatives,

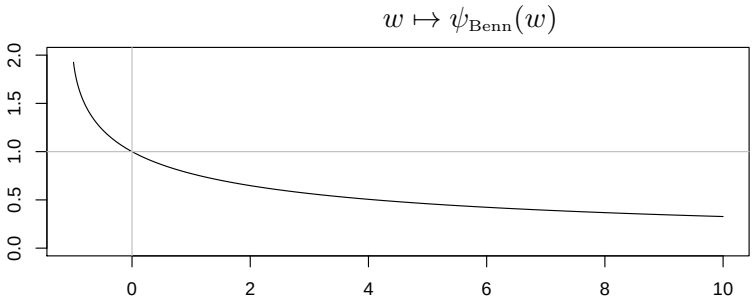
$$\mathbb{f}'(x) = e^x - 1 \quad \text{AND} \quad \mathbb{f}''(x) = e^x$$
$$\mathbb{h}'(x) = \log(1+x) \quad \text{AND} \quad \mathbb{h}''(x) = (1+x)^{-1},$$

show that both functions are non-negative and convex, with each achieving its minimum value of 0 at $x = 0$. From <1> we have representations

\E@ Dele.def <5>
$$\mathbb{f}(x) = \tfrac{1}{2}x^2\mathbb{\Delta}(x) \qquad \text{where } \mathbb{\Delta}(x) := \int_0^1 2(1-s)e^{xs}ds$$

\E@ Bennpsi.def <6>
$$\mathbb{h}(x) = \tfrac{1}{2}x^2\psi_{\text{Benn}}(x) \qquad \text{where } \psi_{\text{Benn}}(x) := \int_0^1 \frac{2(1-s)}{(1+xs)}ds \text{ for } x \geq -1.$$

The function $\mathbb{\Delta}$ is continuous, convex, and strictly increasing. The function ψ_{Benn} is continuous, convex, and strictly decreasing on $[-1,\infty)$.



Moreover, by Jensen's inequality,

\E@ Benn.Bern <7> $$\psi_{\text{Benn}}(x) \geq \hbar''(x/3) = (1 + x/3)^{-1}.$$

Compare with the fact that $\psi_{\text{Benn}}(x)$ actually decreases like $2x^{-1}\log(x)$ as x tends to infinity.

As you will see in Chapter 3, the function ψ_{Benn} lies hidden inside a handful of very useful tail bounds. Inequality <7> will also have some surprising consequences: in Section 2.6, in the context of the Pinsker inequality; and in Chapter 8, to explain the relationship between the Bennett and Bernstein exponential inequalities.

Remarks. The subscript 'Benn' stands for George Bennett, who established a now famous exponential tail bound (Bennett, 1962, equation 8b), although he did not write it using the ψ_{Benn} function. This bound, and its martingale extension, will be described in Chapter 8.

The monotonicity and convexity of $\mathbb{\Delta}$ are easy to check for $x \geq 0$ from the series expansion

$$\mathbb{\Delta}(x) = 2\sum\nolimits_{k\geq 2} x^{k-2}/k!\,.$$

However, even the monotonicity is less obvious when x is negative. Brute force calculation of the derivative,

$$d\mathbb{\Delta}(x)/dx = \frac{(x-2)e^x + x + 2}{x^3/2},$$

didn't help me much when I first encountered $\mathbb{\Delta}$ in the Chow and Teicher (1978, Section 11.1) book. The latest edition of that book did use the shorter proof, almost as above. See also the proof of Freedman (1975, Lemma 3.1).

2.3 Convexity of the log MGF

Calculus::S:MGFconvexity

For a random variable X, the function $M_X(\lambda) = \mathbb{P}\exp(\lambda X)$ for $\lambda \in \mathbb{R}$ is called the moment generating function (MGF) for X. It is always fi-

nite at the origin because $M_X(0) = \mathbb{P}1 = 1$. The expectation is well defined and nonnegative for all real λ, although it might take the value $+\infty$. For example, if X has a Cauchy distribution then $M_X(\lambda) = +\infty$ for all nonzero λ; and if X has a standard exponential distribution then

$$M_X(\lambda) = \int_0^\infty e^{\lambda x - x}\, dx = \begin{cases} (1-\lambda)^{-1} & \text{for } \lambda < 1 \\ \infty & \text{otherwise} \end{cases} .$$

The function M_X inherits convexity from the exponential function. More precisely, the set

$$\text{EPI}(M_X) = \{(\lambda, t) \in \mathbb{R}^2 : t \geq M_X(\lambda)\}$$

is convex and closed as a subset of $\mathbb{R}^2$. The function M_X is a closed, proper ($M_X(\lambda) > -\infty$), convex function, in the sense described by Rockafellar (1970, Section 7). Moreover, the restriction of M_X to the convex set $\mathcal{D}_X := \{\lambda \in \mathbb{R} : M_X(\lambda) < \infty\}$ is continuous and convex in the usual sense. If $\text{int}(\mathcal{D}_X)$, the interior of $\mathcal{D}_X$, is non-empty then M_X is infinitely differentiable with

\E@ MGF.derivs <8> $$M_X^{(k)}(\lambda) = \mathbb{P}\left(X^k e^{\lambda X}\right) \qquad \text{for } k = 1, 2, \ldots \text{ if } \lambda \in \text{int}(\mathcal{D}_X).$$

See Problem [3] for the (easy) proofs.

Remark. The differentiability can fail at the boundary of $\mathcal{D}_X$. For example, if X has the distribution with density $x^{-2}\{x \geq 1\}$ with respect to Lebesgue measure then $M_X(\lambda) < \infty$ iff $\lambda \leq 0$. The left-hand derivative at $\lambda = 0$ is infinite: the ratio $(M_X(0) - M_X(-h))/h$ tends to $+\infty$ as h decreases to zero.

The function $L_X := \log M_X$ inherits from M_X the lower semi-continuity and infinite differentiability on the interior of $\mathcal{D}_X$. More surprisingly, L_X is also a proper, closed convex function. The convexity follows from the Hölder inequality: for nonnegative random variables Y_1 and Y_2: for positive constants α_i with $\alpha_1 + \alpha_2 = 1$,

$$\mathbb{P}\left(Y^{\alpha_1} Y_2^{\alpha_2}\right) \leq (\mathbb{P}Y)^{\alpha_1} \left(\mathbb{P}Y_2\right)^{\alpha_2} .$$

Taking $Y_i = \exp(\lambda_i X)$ we get

$$M_X(\alpha_1\lambda_1 + \alpha_2\lambda_2) \leq M_X(\lambda_1)^{\alpha_1} M_X(\lambda_2)^{\alpha_2}.$$

By taking logs we get

$$L_X(\alpha_1\lambda_1 + \alpha_2\lambda_2) \leq \alpha_1 L_X(\lambda_1) + \alpha_2 L_X(\lambda_2).$$

Remark. If you worry about things like $\log(6\infty) = \log(6) + \log(\infty)$ you could instead note that the restriction of L_X to $\mathcal{D}_X$ is convex in the usual sense, with $L_X(\lambda) = \infty$ outside $\mathcal{D}_X$, which is equivalent to convexity of $\text{EPI}(L_X)$ as a subset of $\mathbb{R}^2$.

Formula <8> leads to interesting expressions for the first two derivatives of L_X on $\text{int}(\mathcal{D}_X)$:

$$L'(\lambda) = \frac{M_X'(\lambda)}{M_X(\lambda)} = \mathbb{P}X \frac{e^{\lambda X}}{M_X(\lambda)}$$
$$L''(\lambda) = \frac{M_X''(\lambda)}{M_X(\lambda)} - \left(\frac{M_X'(\lambda)}{M_X(\lambda)}\right)^2 = \mathbb{P}X^2 \frac{e^{\lambda X}}{M_X(\lambda)} - \left(\mathbb{P}X \frac{e^{\lambda X}}{M_X(\lambda)}\right)^2.$$

If we define a probability measure $\mathbb{P}_\lambda$ by its density, $d\mathbb{P}_\lambda/d\mathbb{P} = e^{\lambda X}/M_X(\lambda)$, then

\E@ LX.derivs <9> $$L'(\lambda) = \mathbb{P}_\lambda X \qquad \text{AND} \qquad L''(\lambda) = \mathbb{P}_\lambda \left(X - \mathbb{P}_\lambda X\right)^2 = \text{var}_\lambda(X) \geq 0.$$

In particular, if $0 \in \text{int}(\mathcal{D}_X)$ then $L'(0) = \mathbb{P}X$, so that Taylor expansion gives

\E@ L.Taylor <10> $$L(\lambda) = \lambda \mathbb{P}X + \tfrac{1}{2}\lambda^2 \text{var}_{\lambda^*}(X) \qquad \text{with } \lambda^* \text{ lying between 0 and } \lambda.$$

I was quite surprised when I first learned how often <9> lurks behind useful facts. Problem [5] presents an example, which I learned from one of my colleagues only last year.

Remark. Bounds on the $\text{var}_{\lambda^*}(X)$, for some hard to calculate λ^* that depends on λ, lead to bounds on L and tail bounds for $X - \mathbb{P}X$. The proof of the Hoeffding inequality in Chapter 7 uses the representation <10> to control the MGF of a bounded random variable.

2.4 Fenchel-Legendre conjugates

Calculus::S:Fenchel

Many probability bounds involve a minimization problem of the form

$$\text{for each } w \in \mathbb{R} \text{ minimize over } \lambda \in J \text{ the function } f(\lambda) - \lambda w,$$

where J is a nonempty convex subset of the real line, and $f : J \to \mathbb{R}$. Of course a convex J is some sort of interval, such as $\mathbb{R}$ itself or $(-\infty, 3]$ or $[-7, 3)$, and so on. The role of J can be ignored if we extend f to a function taking values in $\mathbb{R} \cup \{\infty\}$ by defining $f(\lambda) = +\infty$ for $\lambda \notin J$.

While we are at it, we should allow for the possibility that there might be no minimizing value by defining

\E@ conjugate <11>
$$g(w) = \inf_{\lambda\in\mathbb{R}} \Big(f(\lambda) - \lambda w \Big) \qquad \text{for } w \in \mathbb{R}.$$

Equivalently,

$$-g(w) = \sup_{\lambda\in\mathbb{R}} \Big(\lambda w - f(\lambda) \Big) \qquad \text{for } w \in \mathbb{R}.$$

The function $f^* = -g$ is often called the ***Fenchel-Legendre transform***, or ***convex conjugate***, of f.

Remark. Note that for any function $\ell : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$,

$$\ell(w) + f(\lambda) \geq w\lambda \qquad \text{for all } w, \lambda \in \mathbb{R}$$

if and only if $\ell(w) \geq \mathbb{f}^*(w)$ for all w.

The calculation of the infimum in <11> is easier when f itself is convex. It is then usually easy to crank out g by brute force Calculus by finding where the derivative is positive or negative or zero. In Chapter 3 you will encounter several such cases where $f(\lambda) = \log \mathbb{P} e^{\lambda X}$ for a random variable X.

It is also convenient that g, as an infimum of linear functions of w, always turns out to be concave and f^* is always convex. The convexity can also be seen from consideration of the ***epigraph***:

$$\begin{aligned} \operatorname{epi}(f^*) &= \{(t,w) \in \mathbb{R}^2 : t \geq f^*(w)\} \\ &= \cap_{\lambda\in\mathbb{R}} \{(t,w) \in \mathbb{R}^2 : t - \lambda w \geq -f(\lambda)\}, \end{aligned}$$

an intersection of closed half-spaces. The epigraph is closed as a (nonempty) subset of $\mathbb{R}^2$, not just convex. That makes f^* a closed, proper convex function in the sense used by Rockafellar (1970, Section 7).

Remark. For a proper, convex function with $J = \{\lambda : f(\lambda) < \infty\}$ the closedness property just involves the values of f at the endpoints of J, if any. This subtlety need not concern us.

The general theory is interesting because $f^{**} = (f^*)^*$ happens to be the largest closed, convex function for which $f(w) \geq f^{**}(w)$ for all w (see Rockafellar (1970, pages 103-4). In particular, if f itself is closed and convex then $f^{**} = f$. Such a property plays a central role role in the study of convex duality (Rockafellar, 1974).

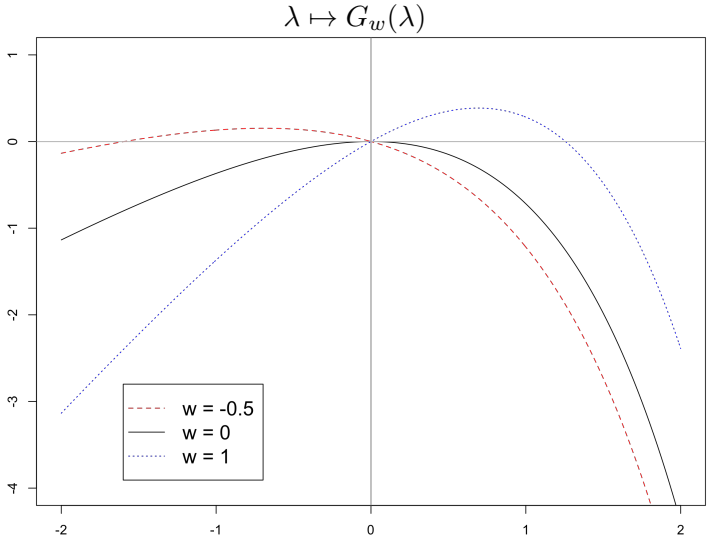
Remark. Unfortunately these wonderful duality facts about f^{**} don't seem to help much for the material in this book. I should perhaps look more carefully at the pages of Boucheron et al. (2000) pointed at by index entries of the form "duality ...".

Calculus::exp <12>

Example. Consider the convex function $\mathbb{f}(\lambda) = e^\lambda - 1 - \lambda$, for $\lambda \in \mathbb{R}$, from Section 2.2. It is convex on the real line with derivative 0 at the origin. For each $w \in \mathbb{R}$ the concave function

$$G_w(\lambda) = \lambda w - \mathbb{f}(\lambda) = \lambda(w+1) - e^\lambda + 1$$

has λ-derivative $G'_w(\lambda) = (1+w) - e^\lambda$.



If $w > -1$ the maximum of G_w is achieved when $G'_w(\lambda) = 0$, that is, at $\lambda = \log(1+w)$. For $w = -1$ we have $G_{-1}(\lambda) = 1 - e^\lambda$, which approaches its supremum of 1 as $\lambda \to \infty$. For $w < -1$, we have $G_w(\lambda) \uparrow +\infty$ as $\lambda \downarrow -\infty$. In summary,

\E@ fbb.star <13>

$$\mathbb{f}^*(w) = \begin{cases} (1+w)\log(1+w) - w & \text{if } w > -1\text{; achieved at } \lambda = \log(1+w) \\ 1 & \text{if } w = -1\text{; approached as } \lambda \to \infty \\ +\infty & \text{if } w < -1\text{; approached as } \lambda \to \infty \end{cases}$$

That is, $\mathbb{f}^* = \mathbb{h}$, as a map from $\mathbb{R}$ to $(-\infty, +\infty]$.

The general duality theory asserts that

$$\mathbb{f}(\lambda) = \mathbb{h}^*(\lambda) = \sup_{w\in\mathbb{R}} w\lambda - \mathbb{h}(w) = \sup_{w\ge -1} w\lambda - \mathbb{h}(w).$$

Let me check.

For $\lambda \in \mathbb{R}$ the supremum defining $\hbar^*(\lambda)$ need only run over $w \geq -1$ because $\lambda w - \hbar(w) = -\infty$ for $w < -1$.

The derivative of the concave function $w \mapsto \lambda w - \hbar(w)$ is $\lambda - \log(1+w)$, which tends to $+\infty$ as $w \to -1$ and $-\infty$ as $w \to \infty$. For the maximizing w we have $\lambda = \log(1+w)$, which gives

$$\hbar^*(\lambda) = \lambda(e^\lambda - 1) - \lambda e^\lambda + (e^\lambda - 1) = e^\lambda - 1 - \lambda.$$

That is, $\hbar^* = \mathbb{f}$, as asserted by the general theory.

2.5 Approximation of factorials

Calculus::S:factorial

Abraham de Moivre (1756, page 244), before deriving his famous approximation to the Binomial distribution, made a comment about a constant B (which we now know is equal to $\sqrt{2\pi}$), which he had approximated by adding up some terms of a slowly converging infinite series. He went on:

> . . . seeing at the same time that what I had done answered my purpose tolerably well, I desisted from proceeding farther till my worthy and learned Friend Mr. *James Stirling*, who had applied himself after me to that inquiry, found that the Quantity B did denote the Square-root of the Circumference of a Circle whose Radius is Unity . . .

Yes, it was the Stirling of the approximation that is now often cited in the exquisitely sharp form: for $n \in \mathbb{N}$,

\E@ Stirling <14>

$$n! = \sqrt{2\pi}\, n^{n+1/2} e^{-n+r_n} \qquad \text{where } b_n := \frac{1}{12n+1} < r_n < \frac{1}{12n} =: a_n.$$

Equivalently, $n! = n^{n+1/2}e^{-n+d_n}$ where

$$d_n := \log\left(n! e^n / n^{n+1/2}\right) = \log(n!) - (n+\tfrac{1}{2})\log n + n$$

and

\E@ sandwich <15>

$$d_n - a_n < C < d_n - b_n \qquad \text{where } C = \log\sqrt{2\pi}.$$

For probabilists, there is a charming connection with the probabilities $p_i = \mathbb{P}\{X_n = i\}$ for $X_n \sim \text{Poisson}(n)$. The maximum occurs at $i = n$,

with $p_n = e^{-n}n^n/n! = n^{-1/2}e^{-d_n}$. Convergence of d_n to $\log\sqrt{2\pi}$ is equivalent to convergence of $\sqrt{n}p_n$ to $1/\sqrt{2\pi}$. The $N(n,n)$ approximation to the POISSON(n) suggests that

$$\begin{aligned} p_n = \mathbb{P}\{n-\tfrac{1}{2} \le X \le n+\tfrac{1}{2}\} &= \mathbb{P}\{-\frac{1}{2\sqrt{n}} \le \frac{X-n}{\sqrt{n}} \le \frac{1}{2\sqrt{n}}\} \\ &\approx \mathbb{P}\{-\frac{1}{2\sqrt{n}} \le Z \le \frac{1}{2\sqrt{n}}\} \approx 1/\sqrt{2\pi n}. \end{aligned}$$

The usual central limit theorem does not quite justify all this handwaving but the basic idea works. See Problem [6] for a rigorous argument.

The proof of <15> is quite beautiful. It is worth your attention. The assertion will follow from the inequalities

\E@ diffs <16>
$$\frac{1}{12n^2+14n+13/12} = b_n - b_{n+1} < d_n - d_{n+1} < a_n - a_{n+1} = \frac{1}{12n(n+1)},$$

which together show that $\{d_n - a_n : n \in \mathbb{N}\}$ is an increasing sequence and $\{d_n - b_n : n \in \mathbb{N}\}$ is a decreasing sequence. It follows that there exists a finite constant C for which $d_n - a_n \uparrow C$ and $d_n - b_n \downarrow C$. Problem [6] shows that $C = \log(\sqrt{2\pi})$, the logarithm of de Moivre's B.

To prove inequality <16> first write everything in terms of $T_n = t_n^{-1} = 2n+1$. By direct calculation

$$\begin{aligned} d_n - d_{n+1} &= (n+1/2)\log\left(\frac{n+1}{n}\right) - 1 \\ &= \tfrac{1}{2}T_n \log\left(\frac{T_n+1}{T_n-1}\right) - 1 = \frac{1}{2t_n}\log\left(\frac{1+t_n}{1-t_n}\right) - 1 \\ &= \frac{t_n^2}{3} + \frac{t_n^4}{5} + \frac{t_n^6}{6} + \dots \qquad \text{by Taylor expansion} \\ &< \frac{t_n^2}{3} + \frac{t_n^4}{3} + \frac{t_n^6}{3} + \dots = \frac{t_n^2}{3}\left(1-t_n^2\right)^{-1} \\ &= \frac{1}{3(T_n^2-1)} = \frac{1}{12n(n+1)} = a_n - a_{n+1}. \end{aligned}$$

Similarly

$$d_n - d_{n+1} > \frac{t_n^2}{3} = \frac{1}{3T_n^2} = \frac{1}{12n^2+12n+3} > b_n - b_{n+1}.$$

The second inequality comes from

$$(12n^2+14n+13/12) - (12n^2+12n+3) = 2n - 23/12 \ge 1/12.$$

Done.

Calculus::binom.coeff <17> **Example.** A lot of asymptotic statistical theory involves the binomial coefficients:

$$\binom{n}{i} = \frac{n!}{i!(n-i)!} \qquad \text{for } i = 0, 1, \ldots, n.$$

To avoid an overabundance of parentheses, for $0 < i < n$ define $\eta = i/n = 1 - \gamma$ and $\epsilon_{i,n} = r_i + r_{n-i} - r_n$. Then

$$\begin{aligned}\binom{n}{i} = \frac{n!}{i!(n-i)!} &= \frac{n^{n+1/2}\exp(\epsilon_{i,n})}{\sqrt{2\pi}(n\eta)^{i+1/2}(n\gamma)^{n-i+1/2}} \qquad \text{by <14>} \\ &= \frac{1}{\sqrt{2\pi n\eta\gamma}} \exp\left(-n\Big[\eta\log\frac{1}{\eta} + \gamma\log\frac{1}{\gamma}\Big] + \epsilon_{i,n}\right).\end{aligned}$$

The $\epsilon_{i,n}$ is usually ignored if both i and $n-i$ are large. The term within square brackets in the exponent is called the ***entropy*** of the $\text{BER}(\eta)$ distribution, a quantity much loved by information theorists (Cover and Thomas, 2012, Example 2.1.1). When multiplied by n it gives the entropy of the $\text{BIN}(n, \eta)$ distribution. □

<18> **Example.** If $X \sim \text{BIN}(n, p)$ and $q = 1 - p$ then, again writing $\eta = 1 - \gamma$ for i/n, we have

$$\begin{aligned}\mathbb{P}\{X = i\} = \binom{n}{i} p^i q^{n-i} &= \binom{n}{i} \exp\left(n\eta\log p + n\gamma\log q\right) \\ &= \frac{1}{\sqrt{2\pi n\eta\gamma}} \exp\left(-n\Big[\eta\log\frac{p}{\eta} + \gamma\log\frac{q}{\gamma}\Big] + \epsilon_{i,n}\right).\end{aligned}$$

The quantity

$$D_{KL}(\eta \,||\, p) := \eta\log\left(\eta/p\right) + \gamma\log\left(\gamma/q\right), \qquad \text{for } q = 1-p \text{ and } \gamma = 1-\eta,$$

is called the ***Kullback-Leibler distance*** (or KL-divergence) between the $\text{BER}(\eta)$ and $\text{BER}(p)$ distributions. Multiplication by n gives the KL-divergence between the $\text{BIN}(n, \eta)$ and the $\text{BIN}(n, p)$ distributions.

The KL-divergence is also closely related to the function ψ_{Benn}, defined as in <6> by $\frac{1}{2}t^2\psi_{\text{Benn}}(t) = \hbar(t) = (1+t)\log(1+t) - t$. If $i = np + x$ then $\eta = p(1 + x/(np))$ and $\gamma = q(1 - x/(nq))$ then $D(\text{BIN}(n,\eta) \,||\, \text{BIN}(n,p))$

equals

$$\begin{aligned} & np\left(1+\frac{x}{np}\right)\log\left(1+\frac{x}{np}\right)+nq\left(1-\frac{x}{nq}\right)\log\left(1-\frac{x}{nq}\right) \\ &= np\times\hbar\left(\frac{x}{np}\right)+nq\times\hbar\left(\frac{-x}{nq}\right)+np\left(\frac{x}{np}\right)+nq\left(\frac{-x}{nq}\right) \\ &= \frac{x^2 g(x,n,p)}{2npq} \qquad \text{where } g(x,n,p):=q\psi_{\text{Benn}}\left(\frac{x}{np}\right)+p\psi_{\text{Benn}}\left(\frac{-x}{nq}\right). \end{aligned}$$

In summary, for $0<i=np+x<n$ and $X\sim \text{BIN}(n,p)$ we have

\E@ Bin.rep <19>
$$\mathbb{P}\{X=i\}=\frac{1}{\sqrt{2\pi npq}}\exp\left(-\mathfrak{Q}(x)\right)\times\Lambda(x,n,p)$$

where

$$\begin{aligned} \mathfrak{Q}(x) &= np\times\hbar\left(\frac{x}{np}\right)+nq\times\hbar\left(\frac{-x}{nq}\right)=\frac{x^2 g(x,n,p)}{2npq} \\ g(x,n,p) &= q\psi_{\text{Benn}}\left(\frac{x}{np}\right)+p\psi_{\text{Benn}}\left(\frac{-x}{nq}\right) \\ \log\Lambda(x,n,p) &:= r_{np+x}+r_{nq-x}-r_n-\tfrac{1}{2}\log\left[\left(1+x/(np)\right)\left(1-x/(nq)\right)\right]. \end{aligned}$$

The constraint $0<i<n$ is only needed to ensure that there are no divisions by 0 in the definition of Λ.

If $\min(np,nq)$ is large and $x/\min(np,nq)$ is small then $\Lambda(x,n,p)\approx 1$ and $g(x,n,p)\approx 1$, so that

$$\mathbb{P}\{X=i\}\approx\frac{1}{\sqrt{2\pi npq}}\exp\left(-\frac{x^2}{2npq}\right).$$

This result should remind you of the normal approximation to the $\text{BIN}(n,p)$, but not just as an approximation for tail probabilities. A more careful handling of the errors of approximation would lead to a so-called ***local limit theorem*** for the Binomial distribution (Petrov, 1975, Chapter 7).

If you don't like great gobs of algebra you could safely skip the rest of this Example. I am just about to make a small point about a refinement of the normal approximation to the Binomial, with another advertisement for the methods from Section 2.2.

The appearance of the convex function ψ_{Benn} in the definition of $g(x,n,p)$ tell us something surprising about departures from normality. By convexity,

\E@ Bennpsi.convex <20>
$$g(x,n,p)\geq\psi_{\text{Benn}}\left(q\frac{x}{np}+p\frac{-x}{nq}\right)=\psi_{\text{Benn}}\left(\frac{(q-p)x}{npq}\right).$$

If $p > 1/2$ then $g(x, n, p) > 1$ for $x > 0$. We get something smaller than the normal exponent $-x^2/(2npq)$ in the right tail beyond the mean np. This effect can be attributed to the skewness, $\mathbb{P}(X - np)^3/\text{var}^{3/2}(X) = npq(q-p)/(npq)^{3/2}$, which is negative if $p > 1/2$. To a first approximation, the $\text{Bin}(n, p)$ is symmetric around its mean np. To a finer approximation, if $p > 1/2$ then the probabilities are slightly smaller in the left tail than in the right tail.

By the principle of no free lunches, if skewness gives us a nicer right tail then it will probably give us an uglier left tail. That is in fact the case but it cannot be deduced directly from inequality <20>. Instead we need to identify some higher order terms in the series expansion of $\mathfrak{Q}(x)$. From

$$\begin{aligned} \mathfrak{h}'(x) &= \log(1+x) & \mathfrak{h}''(x) &= 1/(1+x) \\ \mathfrak{h}^{(3)}(x) &= -1/(1+x)^2 & \mathfrak{h}^{(4)}(x) &= 2/(1+x)^3 \end{aligned}$$

and Problem [1] we get

$$\mathfrak{h}(x) = \frac{x^2}{2} - \frac{x^3}{6} + \frac{x^4}{3}\int_0^1 \frac{(1-s)^3}{(1+sx)^3}\,ds \qquad \text{for } x > -1$$

and

$$\begin{aligned} \mathfrak{Q}(x) &= nq\mathfrak{h}(x/np) + np\mathfrak{h}(-x/nq) \\ &= np\left(\frac{x^2}{2(np)^2} - \frac{x^3}{6(np)^3} + \frac{x^4}{3(np)^4}\int_0^1 \frac{(1-s)^3}{(1+sx/np)^3}ds\right) \\ &\quad + nq\left(\frac{x^2}{2(nq)^2} + \frac{x^3}{6(nq)^3} + \frac{x^4}{3(nq)^4}\int_0^1 \frac{(1-s)^3}{(1-sx/nq)^3}ds\right) \\ &= \frac{x^2}{2npq} + \frac{x^3(q-p)}{6n^2p^2q^2} \\ &\qquad + \frac{x^4}{3}\int_0^1 (1-s)^3\left[(np+sx)^{-3} + (nq-sx)^{-3}\right]\,ds. \end{aligned}$$

The pattern is clearer if we consider terms $i = np + x$ with $x = y\sqrt{npq}$, that is, if we scale by the standard error $\sigma = \sqrt{npq}$ for X:

$$\begin{aligned} \mathfrak{Q}(y\sigma) &= \frac{y^2}{2} + \frac{y^3(q-p)}{6\sqrt{npq}} \\ &\qquad + \frac{y^4p^2q^2}{3n}\int_0^1 (1-s)^3\left[(p+sy\sigma/n)^{-3} + (q-sy\sigma/n)^{-3}\right]\,ds. \end{aligned}$$

The y^2 term contributes the normal approximation; the y^3 term contributes a skewness perturbation of order $n^{-1/2}$; and the y^4 term scoops up the leftovers, which are of order n^{-1}. If you like this kind of thing then you will love Petrov (1975, Chapter 6). □

Calculus::VC <21> **Example.** For many purposes a somewhat crude upper bound for $\binom{n}{i}$ suffices. For example, if $0 < i \le d < n/2$,

$$\binom{n}{i} = \frac{n(n-1)\dots(n-i+1)}{i!} < \left(\frac{n}{d}\right)^i \frac{d^i}{i!} < \left(\frac{n}{d}\right)^d \frac{d^i}{i!} < \left(\frac{en}{d}\right)^d.$$

You will see these bound used in a lot of statistical papers involving high-dimensional (dimension d) data.

You will also see a related expression in the ultra-famous Vapnik & Červonenkis theory:

$$\sum\nolimits_{0\le i\le d} \binom{n}{i} \le \sum\nolimits_{0\le i\le d} \left(\frac{n}{d}\right)^d \frac{d^i}{i!} < \left(\frac{en}{d}\right)^d.$$

See Section 13.2 for more of that exciting story. □

2.6 Distances between probability measures

Calculus::S:divergence

Here is another example of ψ_{Benn} popping up in unexpected places.

Modern statistical theory makes extensive use of various measures of 'distance' between probability measures P and Q defined on the same sigma-field. To keep the following discussion simple I'll assume that both probabilities are defined on the same finite set $\mathcal{X}$ with $Q\{x\} > 0$ for each $x \in \mathcal{X}$. Define

$$\rho(x) = \frac{P\{x\}}{Q\{x\}} \qquad \text{for } x \in \mathcal{X}.$$

For each real-valued function g on $\mathcal{X}$ think of its Q expectation as the linear functional

$$Qg = Qg(x) := \sum\nolimits_{x\in\mathcal{X}} Q\{x\}g(x).$$

Remark. The quantity $\rho(x)$ is sometimes called the ***likelihood ratio***. If you desire greater generality, feel free to extend the arguments to general sigma-fields with P absolutely continuous with respect to Q with density (Radon-Nikodym derivative) $\rho = dP/dQ$. You will also need to worry about integrability. I would even forgive you for adopting the hideous notation $\mathbb{E}_Q g$ for Qg, but the even more hideous $\mathbb{E}_{x\sim Q} g(x)$ is beyond the pale, in my humble opinion.

For each convex function f on $\mathbb{R}^+$ for which $f(1) = 0$, the f-divergence is defined as $D_f(P \mid\mid Q) := Qf(\rho(x))$. Three particularly popular special cases correspond to the functions $f(x) = |x-1|$ (the L^1 distance), and $f(x) = x \log x$ (the Kullback-Leibler divergence) and $f(x) = (x-1)^2$ (the χ^2-divergence):

$$\begin{aligned} \|P - Q\|_1 &:= Q|\rho(x) - 1| = \sum\nolimits_x |P\{x\} - Q\{x\}| \\ D_{KL}(P \mid\mid Q) &:= Q\rho(x)\log(\rho(x)) = \sum\nolimits_x P\{x\} \log\left(P\{x\}/Q\{x\}\right) \\ D_{\chi^2}(P \mid\mid Q) &:= Q\left(\rho(x) - 1\right)^2 = \sum\nolimits_x \frac{(P\{x\} - Q\{x\})^2}{Q\{x\}} \end{aligned}$$

Of these, only the L^1 distance is a metric.

The relationship between the three divergences becomes clearer when they are written as integrals involving $\delta(x) = \rho(x) - 1$:

$$\begin{aligned} \|P - Q\|_1 &:= Q|\delta| \\ D_{KL}(P \mid\mid Q) &:= Q\mathbb{h}(\delta) = \tfrac{1}{2} Q\left(\delta^2 \psi_{\text{Benn}}(\delta)\right) \\ D_{\chi^2}(P \mid\mid Q) &:= Q\delta^2 \end{aligned}$$

As before, $\mathbb{h}(x) = (1+x)\log(1+x) - x$. The extra $-x$ does not change the integral because $Q\delta = Q\rho - 1 = 1 - 1$.

The divergences are related by the inequalities

$$D_{\chi^2}(P \mid\mid Q) \geq D_{KL}(P \mid\mid Q) \geq \tfrac{1}{2}\|P - Q\|_1^2 .$$

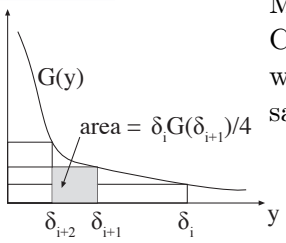
I don't know if anyone has claimed parentage for the first inequality, which follows from the fact that $\psi_{\text{Benn}}(\delta) \leq 2$. The second inequality, which is usually called ***Pinsker's inequality***, follows from Cauchy-Schwarz:

$$\begin{aligned} \tfrac{1}{2}\left(Q|\delta|\right)^2 &= \tfrac{1}{2} Q\left(\frac{|\delta|}{\sqrt{1+\delta/3}}\sqrt{1+\delta/3}\right)^2 \\ &\leq \tfrac{1}{2} Q\left(\frac{\delta^2}{1+\delta/3}\right) Q(1+\delta/3) \qquad \text{by Cauchy-Schwarz} \\ &\leq \tfrac{1}{2} Q\left(\delta^2 \psi_{\text{Benn}}(\delta)\right) \qquad \text{by <7> and } Q\delta = 0. \end{aligned}$$

I learned of this neat argument from Kemperman (1969).

2.7 Bounding sums by integrals

Calculus::S:sum.to.integral



Many approximation schemes (particularly those that will be discussed in Chapter 9) bound useful quantities by sums of the form $\sum_{i=0}^{k} \delta_i G(\delta_{i+1})$, with $G(\cdot)$ a decreasing function on $\mathbb{R}^+$. If the δ_i's decrease geometrically, say $\delta_i = \delta_0/2^i$, then the sum can be bounded above by an integral,

$$\sum_{i=0}^{k} \delta_i G(\delta_{i+1}) \le 4 \int_{\delta_{k+2}}^{\delta_1} G(r)\, dr \le 4 \int_0^{\delta_1} G(r)\, dr.$$

If G is integrable the replacement of the lower terminal by 0 costs little.

In the early empirical process literature, some authors chose the δ_i's to make the $G(\delta_i)$'s increase geometrically, for similar reasons. (Use horizontal slices.)

In delicate situations, sometimes the $\{\delta_i\}$ sequence needs to chosen in a more cunning way. See Section 9.6, for example.

2.8 Problems

Calculus::S:Problems

Calculus::P:Taylor.Rk

[1] Suppose g is a real-valued function, defined at least on an interval J of the real line that contains both 0 and a point x, that is $(k+1)$-times continuously differentiable. For x in the interior of J show that

$$\begin{aligned} R_k(x,g) &:= g(x) - \left[g(0) + xg'(0) + \cdots + x^k g^{(k)}(0)/(k)!\right] \\ &= x^{k+1} \int \cdots \int \{0 \le t_{k+1} \le t_k \le \cdots \le t_1 \le 1\} g^{(k)}(t_{k+1}x)\, dt_{k+1} \ldots dt_1 \\ &= x^{k+1} \int_0^1 \frac{(1-s)^k}{k!} g^{(k+1)}(sx)\, ds. \end{aligned}$$

Calculus::P:eix

[2] By splitting into real and imaginary parts, show that

$$e^{ix} - \left(1 + ix + (ix)^2/2! + \cdots + (ix)^k/k!\right) = x^{k+1} \int_0^1 \frac{(1-s)^k}{k!} i^{k+1} e^{isx}\, ds.$$

for all real x. Deduce that

$$\left| e^{ix} - \left(1 + ix + (ix)^2/2! + \cdots + (ix)^k/k!\right) \right| \le \frac{|x|^{k+1}}{(k+1)!},$$

an inequality is useful for arguments involving Fourier transforms.

Calculus::P:smooth.MGF

[3] Suppose $M_X(\lambda) = \mathbb{P}e^{\lambda X}$ for $\lambda \in \mathbb{R}$ and $\mathcal{D}_X = \{\lambda \in \mathbb{R} : M_X(\lambda) < \infty\}$.

(i) Show that $M_X(\lambda)$ is lower semi-continuous at each point of $\mathbb{R}$. Equivalently, $\text{EPI}(M_X)$ is a closed subset of $\mathbb{R}^2$. Hint: If $\lambda_n \to \lambda \in \mathbb{R}$, Fatou's lemma tells you something about $\liminf_n \mathbb{P}e^{\lambda_n X}$. Argue similarly if $(t_n, \lambda_n) \in \text{EPI}(M_X)$ and $(t_n, \lambda_n) \to (t, \lambda) \in \mathbb{R}^2$. (As Rockafellar (1970, page 52) argued, lower semi-continuity of M_X is equivalent to $\text{EPI}(M_X)$ being closed as a subset of $\mathbb{R}^2$.)

(ii) Show that the restriction of M_X to $\mathcal{D}_X$ is continuous as a function on $\mathcal{D}_X$. Hint: Suppose $[\lambda, \gamma] \subset \mathcal{D}_X$ and $\lambda_n = \alpha_n\lambda + (1-\alpha_n)\gamma$. Convexity of M_X gives

$$M_X(\lambda_n) \le \alpha_n M_X(\lambda) + (1-\alpha_n)M_X(\gamma). \qquad \text{for } 0 \le t \le 1.$$

Take the lim sup as α_n tends to 0 to deduce that $\limsup M_X(\lambda_n) \le M_X(\lambda)$, which together with lower semi-continuity proves continuity from the right at λ. Argue similarly if $\alpha_n \to 1$ for continuity from the left at γ.

(iii) Suppose λ is an interior point of $\mathcal{D}_X$. Choose $\delta > 0$ so that $[\lambda - 2\delta, \lambda + 2\delta] \subset \mathcal{D}_X$. For $|h| < \delta$, establish the domination bound

$$\begin{aligned} |e^{(\lambda+h)X} - e^{\lambda X}|/|h| &\le \left| \int_0^1 X e^{\lambda X} e^{shX} ds \right| \\ &\le e^{\lambda X} \left(e^{\delta X} + e^{-\delta X} \right)^2 / \delta, \end{aligned}$$

which is integrable. Deduce via a Dominated Convergence argument that $M'(\lambda) = \mathbb{P}(Xe^{\lambda X})$.

(iv) Argue similarly to show that M_X is infinitely differentiable on the interior of $\mathcal{D}_X$, with $M_X^{(k)}(\lambda) = \mathbb{P}\left(X^k e^{\lambda X}\right)$.

Calculus::P:log.gamma

[4] The gamma function is defined for $\alpha > 0$ by $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$. Adapt either of the methods from Section 2.3 to prove that $\log\Gamma(\cdot)$ is a convex function.

Calculus::P:condit.mean

[5] (See Wu and Zhou (2019) for an example of how the results from this Problem can be used.) Suppose $X \sim P$ and $Z \sim \mathcal{N} := N(0,1)$ are independent. Define $U = \alpha X + \beta Z$ for positive constants α and β. Define $h(u) = \mathbb{P}(X \mid U = u)$. Follow these steps to shows that

(a) h is an increasing function

(b) If P is symmetric around 0 then h is an odd function.

(i) Explain why the result is trivial if either of α or β is zero. Deduce that, without loss of generality, we may suppose $\alpha = \beta = 1$.

(ii) For suitably integrable f, show that

$$\mathbb{P}f(X, X+Z) = P^x \int_{\mathbb{R}} f(x,u)m(x,u)\phi(u)\,du$$

where $m(x,u) = \phi(x-u)/\phi(u) = \exp(xu - x^2/2)$. Deduce that: the joint distribution $\mathbb{M}$ for (X,U) has density m with respect to $P \otimes \mathcal{N}$; the marginal distribution (with respect to $\mathcal{N}$) for U is $q(u) = P^x m(x,u)$; and the conditional density (with respect to P) for X given $U = u$ is $m(x \mid u) = m(x,u)/q(u)$.

(iii) Assuming the usual conditions for differentiating inside P integrals, deduce that $u \mapsto \log q(u)$ is convex, so that its derivative, $q'(u)/q(u)$, is an increasing function.

(iv) Use the fact that $\partial m(x,u) = x m(x,u)$ to show that

$$h(u) = P^x x m(x,u)/q(u) = P^x \frac{\partial m(x,u)}{\partial u}/q(u) = q'(u)/q(u),$$

an increasing function.

(v) Note that $m(-x,-u) = m(x,u)$ for all x and u. For symmetric P deduce that

$$q(-u) = P^x m(x,-u) = P^x m(-x,-u) = q(u)$$

and

$$q(-u)h(-u) = q(u)P^x x m(x,-u) = q(u)P^x(-x)m(-x,-u) = -q(u)h(u),$$

which implies that h is an odd function.

Calculus::P:Stirling

[6] Section 2.5 noted the similarity between p_n, the maximum of the probabilities $p_i = \mathbb{P}\{X = i\}$ for $X \sim \text{Poisson}(n)$, and the Stirling formula, with the comment that $C = \log(\sqrt{2\pi})$ if and only if $\sqrt{2\pi n}p_n \to 1$ as $n \to \infty$. Here is one way to prove the convergence property for the Poisson.

In what follows let $K = K_n$ be a positive integer of order $o(n^{2/3})$ for which $K_n/\sqrt{n} \to \infty$. Let $\mathcal{K}$ denote the set of integers j with $-K \le j \le K-1$.

(i) If f is Lipschitz function on the real line and δ is positive show that

$$\left|\delta f(i\delta) - \int_{i\delta}^{(i+1)\delta} f(x)\,dx\right| \le \|f\|_{\text{Lip}}\,\delta^2.$$

Deduce that

$$\left|\sum_{i\in\mathcal{K}} \delta f(i\delta) - \int_{-K\delta}^{K\delta} f(x)\,dx\right| \le 2K\,\|f\|_{\mathrm{Lip}}\,\delta^2.$$

In particular, for $f(x) = e^{-x^2/2}$ and $\delta = n^{-1/2}$ deduce that

$$n^{-1/2}\sum_{i\in\mathcal{K}} e^{-i^2/(2n)} = \int_{-K/\sqrt{n}}^{K/\sqrt{n}} e^{-x^2/2}dx + O(K/n) = \sqrt{2\pi} + o(1).$$

(ii) For $i \in \mathbb{N}$ with $i/n \le 1/2$ show that

$$\log\frac{p_{n+i}}{p_n} = \log\frac{n^i}{\prod_{j=1}^{i}(n+j)} = -\sum_{j=1}^{i}\log\left(1+j/n\right) = -\frac{i^2}{2n} + R_i$$

and $\log\left(p_{n-i}/p_n\right) = -i^2/(2n) + R_{-i}$, where, for some universal constant C_2,

$$\max\left(|R_i|, |R_{-i}|\right) \le C_2\left(i/n + i^3/n^2\right).$$

Hint: $\log(1+x) = x + r(x)$ with $|r(x)| \le x^2$ for $|x| \le 1/2$.

(iii) For $i \in \mathcal{K}$ show that

$$p_{n+i} = p_n \exp\left(-i^2/(2n)\right)(1+\epsilon_i)$$

where $\max_{i\in\mathcal{K}}|\epsilon_i| = O(K/n + K^3/n^2) = o(1)$.

(iv) Show that

$$\begin{aligned}\mathbb{P}\{X - n \in \mathcal{K}\} &= \sum_{i\in\mathcal{K}} p_{n+i} \\ &= n^{1/2}p_n\left(n^{-1/2}\sum_{i\in\mathcal{K}} e^{-i^2/(2n)}\right)(1+o(1)) \\ &= n^{1/2}p_n\left(\sqrt{2\pi}+o(1)\right).\end{aligned}$$

(v) Conclude that $\sqrt{2\pi n}p_n \to 1$ as $n \to \infty$.

Calculus::P:crude.Stirling

[7] For $k \in \mathbb{N}$ prove that $e^k \ge k^k/k! \ge 1$. Deduce that $(k/e)^k \le k! \le k^k$ and $k/e \le (k!)^{1/k} \le k$. Compare with the sharper bounds given by Stirling's formula: $(k!)^{1/k} = \rho_k k/e$, where $\rho_k = \left(\sqrt{2\pi k}e^{r_k}\right)^{1/k}$ is a decreasing function of k that converges to 1 as $k \to \infty$.

k	1	2	3	4	5	6	7	8	9	10
ρ_k	2.718	1.922	1.646	1.504	1.416	1.356	1.313	1.279	1.253	1.231

2.9 Notes

Calculus::S:notes

See Stigler (1986, pages 70-76) for an illuminating discussion of the interaction between de Moivre and Stirling. I learned Stirling's formula, as in <14>, from Feller (1968, page 52). He cited Robbins (1955), who commented that the only novelty in his proof was in the derivation of the inequality $b_n < r_n$. He acknowledged that the standard parts of his argument were taken from Darmois (1928, pages 315-317). Robbins also commented that the "editor has pointed out that the inequalities ... permit the following brief proof", which was a "modification of that attributed to Cesàro by A. Fisher, *Mathematical theory of probabilities*, New York, 1936, pp. 93-95." (The Cesàro book was published in 1884. Even though I cannot read Italian, I feel the proof on page 270 is very similar to the one presented by Feller.) The 1955 Mathematical Monthly listed Carl B. Allendoerfer as the editor. As they say, there is nothing new under the sun.

Feller (1968, footnote to page 53) cited Feller (1967) for the identification of the limit of $\{e^{d_n}\}$ as $\sqrt{2\pi}$. In another footnote Feller (1968, Section VII.2) recovered the $\sqrt{2\pi}$ from a local limit theorem for the $\text{Bin}(n, 1/2)$. The argument in Problem [6] is essentially that of de Moivre (1756), applied to the Poisson instead of the Binomial. See also Pitman (1993, Section 2.3) for a beautiful exposition of de Moivre's method. At the end of Section 3.6 of the same textbook, Pitman noted that de Moivre's method also works with the hypergeometric distribution. It can be also be applied to other distributions P on the integers for which the ratios $P\{i+1\}/P\{i\}$ have a tractable form. Similar ideas also play a role in the the study of log-concave discrete distributions and even in the Chen-Stein theory of approximation (Barbour, Holst, and Janson, 1992, Section 9.2).

Initially I included Section 2.4 mostly because I feared I had missed some subtle point made by Boucheron, Lugosi, and Massart (2013, Sections 2.2), who seemed to be suggesting that the Fenchel-Legendre transform plays a role far more important than just as a name for a minimization procedure. To me it didn't seem very helpful just to know that a particular minimization of a convex function was an example of a technique named after famous mathematicians. Eventually my opinion changed when it became clearer to me that Fenchel-Legendre conjugates were lurking in the background for many useful ideas. See, for example, the discussion of Young's inequality in Section 5.4. The general theory also extends easily to functions on $\mathbb{R}^d$, where it plays an important role in the study of large deviations for sums of independent random vectors—see Dembo and

Zeitouni (1998, Chapter 2).

There is currently an archive of Fenchel's papers at

`http://web.math.ku.dk/arkivet/fenchel/wfenpapr.htm`

There is also a most informative note ("Werner Fenchel, a pioneer in convexity theory and a migrant scientist" by Christer Oscar Kiselman) about Fenchel and duality at

`http://www.cb.uu.se/~kiselman/bibliography.html`

The note is listed on the web site as having appeared in *Normat. Nordisk matematisk tidskrift* 61, No. 2-4, 133-152, but I have only seen the online version.

References

BarbourHolstJanson92 — Barbour, A. D., L. Holst, and S. Janson (1992). *Poisson Approximation*. Oxford University Press.

Bennett62jasa — Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association 57*, 33–45.

cheronLugosimassart2000ras — Boucheron, S., G. Lugosi, and P. Massart (2000). A sharp concentration inequality with applications. *Random Structures and Algorithms 16*, 277–292.

BLM2013Concentration — Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

ChowTeicher78book — Chow, Y. S. and H. Teicher (1978). *Probability Theory: Independence, Interchangeability, Martingales*. New York: Springer.

CoverThomas2012 — Cover, T. M. and J. A. Thomas (2012). *Elements of Information Theory*. John Wiley & Sons.

Darmois1928 — Darmois, G. (1928). *Statistique Mathématique*. Doin & cie.

deMoivre1756 — de Moivre, A. (1756). *The Doctrine of Chances: or, A Method of Calculating the Probabilities of Events in Play*. New York: Chelsea. Third edition (fuller, clearer, and more correct than the former), reprinted in 1967. First edition 1718.

DemboZeitouni98 Dembo, A. and O. Zeitouni (1998). *Large Deviations Techniques and Applications* (2 ed.). Springer-Verlag.

Dudley73gauss Dudley, R. M. (1973). Sample functions of the Gaussian process. *Annals of Probability 1*, 66–103.

Feller1967AMM Feller, W. (1967). A direct proof of Stirling's formula. *The American Mathematical Monthly 74*(10), 1223–1225.

Feller1 Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (third ed.), Volume 1. New York: Wiley.

Freedman1975AnnProb Freedman, D. A. (1975). On tail probabilities for martingales. *Annals of Probability 3*(1), 100–118.

kemperman69 Kemperman, J. H. B. (1969). On the optimum rate of transmitting information. In *Probability and Information Theory*. Springer-Verlag. Lecture Notes in Mathematics, 89, pages 126–169.

Petrov1975 Petrov, V. V. (1975). *Sums of Independent Random Variables*. Springer-Verlag. Enlish translation from 1972 Russian edition.

Pitman1993Prob Pitman, J. (1993). *Probability*. Springer.

Robbins1955AMM Robbins, H. (1955). A remark on Stirling's formula. *The American Mathematical Monthly 62*(1), 26–29.

Rockafellar70book Rockafellar, R. T. (1970). *Convex Analysis*. Princeton, New Jersey: Princeton Univ. Press.

Rockafellar1974conjugate Rockafellar, R. T. (1974). *Conjugate Duality and Optimization*, Volume 16 of *Regional Conference Series in Applied Mathematics*. Siam.

Stigler86book Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, Massachusetts: Harvard University Press.

WuZhou2019EM Wu, Y. and H. H. Zhou (2019). Randomly initialized em algorithm for two-component gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. Technical report.