Printed: 28 September 2024 at 17:29

Chapter 3

The moment generating function method

MGF::MGF

SECTION 3.1 *introduces the MGF (moment generating function) method for bounding tail probabilities.*

SECTION 3.3 *illustrates the MGF method for the simplest case, the normal distribution. The normal is the prototype for the subgaussian family of distributions, which will be discussed in Chapter 7.*

*SECTION 3.4 *derives some global properties for the standard normal hazard function $\phi(x)/\bar{\Phi}(x)$. The value of the global approach is illustrated by a brief discussion of Charles Stein's miraculous method for developing normal approximations.*

SECTION 3.5 *derives tail bounds for the* POISSON *distribution. The omnipresent convex function $\psi_{\text{benn}}()$ puts in an appearance. The Poisson is the prototype for Bennett's inequalities, which will be derived in Chapter 8.*

SECTION 3.6 *establishes tail bounds for the gamma distribution, the prototype for Bernstein's inequalities, which will be derived in Chapter 8.*

SECTION 3.7 *establishes very good bounds for the tails of the* BINOMIAL *distribution, which look a lot like a fancier version of the bounds for the* POISSON. *These bounds also work for* POISSON-BINOMIAL *distributions and other sums of independent random variables taking values in $[0, 1]$. Both results follow via the* JENSEN *inequality from the convexity of the exponential function.*

SECTION 3.8 *shows that the tail bounds derived by the MGF method for the hypergeometric distribution (sampling without replacement) are smaller than the bounds for the corresponding* BINOMIAL*(sampling with replacement).*

3.1 Tail bounds from the moment generating function

MGF::S:method

Much modern statistical theory relies on a handful of probabilistic inequalities, often in the form of bounds on tail probabilites or concentration inequalities. This Chapter introduces one of the main methods for establishing such

bounds. By way of illustration, the method is applied to derive bounds for several well studied cases, which provide the prototypes for a handful of very useful tail bounds.

The method uses the MGF, $M_X(\lambda) := \mathbb{P}e^{\lambda X} = e^{L_X(\lambda)}$, to get upper bounds for $\mathbb{P}\{X \geq x\}$. Remember from Section 2.3 that L_X is infinitely differentiable and convex on the set $\{\lambda \in \mathbb{R} : M_X(\lambda) < \infty\}$.

From the fact that the exp() function is everywhere nonnegative and $\exp(\lambda(X - x)) \geq 1$ when $X \geq x$ and $\lambda \geq 0$ we have

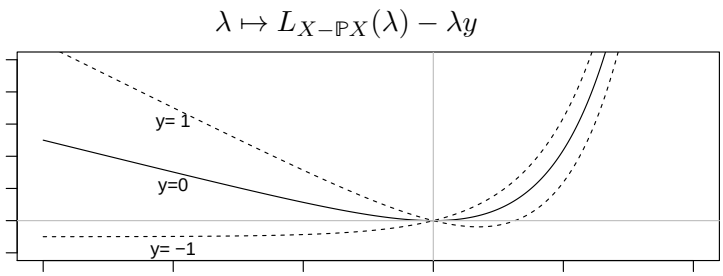
\E@ mgf.upper.tail <1>
$$\mathbb{P}\{X \geq x\} \leq \inf_{\lambda \geq 0} \mathbb{P}e^{\lambda(X-x)} = \inf_{\lambda \geq 0} e^{-\lambda x} M_X(\lambda) = \exp\left(\inf_{\lambda \geq 0}\Big(L_X(\lambda) - \lambda x\Big)\right).$$

Similarly, $\exp(\lambda(X - x)) \geq 1$ if $X \leq x$ and $\lambda \leq 0$, so that

\E@ mgf.lower.tail <2>
$$\mathbb{P}\{X \leq x\} \leq \inf_{\lambda \leq 0} \mathbb{P}e^{\lambda(X-x)} = \exp\left(\inf_{\lambda \leq 0}\Big(L_X(\lambda) - \lambda x\Big)\right).$$

Remark. This inequality can also be derived from <1> applied to bound $\mathbb{P}\{-X \geq -x\}$ by $\inf_{\lambda \geq 0} e^{\lambda(-x)} M_{-X}(\lambda) = \inf_{\lambda \leq 0} e^{-\lambda x} M_X(\lambda)$.

The analysis is simplified if we assume that X has been centered to have $\mathbb{P}X = 0$ and L_X is finite in a neighborhood of the origin, for then $\dot{L}_X(0) = \mathbb{P}X = 0$ and the convex function L_X is minimized at the origin. Equivalently, we can just replace X by $X - \mathbb{P}X$.



Remark. The picture actually shows the case where $X \sim \text{POISSON}(1)$.

The multiplication of $M_{X-\mathbb{P}X}$ by $e^{-\lambda y}$ tilts the convex function $L_{X-\mathbb{P}X}$ by $-\lambda y$, which ensures that $L_{X-\mathbb{P}X}(\lambda) - \lambda y$ achieves its global minimium on the half line $\{\lambda \in \mathbb{R} : \lambda \geq 0\}$ if $y > 0$ and on the half line $\{\lambda \in \mathbb{R} : \lambda \leq 0\}$ if $y < 0$. For the purposes of <1> and <2> we no longer have to consciously think about the sign of y; the infimum in both cases reduces to minimization over the whole real line and everything can be done by brute force calculus. In short,

\E@ LX.upper <3>
$$\mathbb{P}\{X - \mathbb{P}X \geq x\} \leq \exp\left(-\Lambda(x)\right) \qquad \text{for } x \geq 0,$$

\E@ LX.lower <4>
$$\mathbb{P}\{X - \mathbb{P}X \leq x\} \leq \exp\left(-\Lambda(x)\right) \qquad \text{for } x \leq 0,$$

where, for all $y \in \mathbb{R}$,

$$\Lambda(y) := -\inf_{\lambda\in\mathbb{R}} \left(L_{X-\mathbb{P}X}(\lambda) - \lambda y\right) = \sup_{\lambda\in\mathbb{R}} \left(\lambda y - L_{X-\mathbb{P}X}(\lambda)\right).$$

Effectively, we need only search for the global solution to $\dot{L}_{-\mathbb{P}X}(\lambda) = y + \mathbb{P}X$ to determine Λ, except in those pesky cases where the infimum of the convex function $\lambda \mapsto L_{X-\mathbb{P}X}(\lambda) - \lambda y$ is approached as λ tends to $\pm\infty$.

Remark. The nonnegativity of Λ comes from the zero contribution at $\lambda = 0$. The second expression for $\Lambda(y)$ identifies it as $L^*_{X-\mathbb{P}X}(y)$, the FENCHEL-LEGENDRE conjugate of the convex function $L_{X-\mathbb{P}X}$. Exciting as the recognition of this conjugate in a probability bound might be, it does not seem to help much in the actual calculation for a given X. Everything comes down to an exercise in calculus and convexity, which can be worked through without any knowledge of the material in Section 2.4.

*3.2 Behavior of the tail bound near the origin

MGF::S:local

Even though the tail bounds are not particularly useful for small t it is illuminating to see how the moments of $X - \mathbb{P}X$ affect Λ when its MGF is finite in a neighborhood of the origin.

The MGF gets its name from the coefficients in its power series expansion

$$M_{X-\mathbb{P}X}(\lambda) = \mathbb{P}e^{\lambda(X-\mathbb{P}X)} = 1 + \sum_{k\in\mathbb{N}} \mu_k \lambda^k/k! \qquad \text{where } \mu_k = \mathbb{P}\left(X - \mathbb{P}X\right)^k.$$

The quantities μ_k is often called the ***k*th central moment** to distinguish it from $\mathbb{P}X^k$. Note that $\mu_1 = 0$ and $\mu_2 = \operatorname{var}(X)$. The quantity $\mu_3/\mu_2^{3/2} = \mathbb{P}(X - \mathbb{P}X)^3/\operatorname{var}(X)^{3/2}$ is called the ***skewness*** of the distribution.

The function $L_{X-\mathbb{P}X} = \log M_{X-\mathbb{P}X}$ also has a power series expansion,

$$L_{X-\mathbb{P}X}(\lambda) = \sum_{k\in\mathbb{N}} \kappa_k \lambda^k/k! \qquad .$$

The coefficients κ_k are called the (central?) ***cumulants*** and $L_{X-\mathbb{P}X}$, not surprisingly, is the ***cumulant generating function***. Note that there is no κ_0, because $L_{X-\mathbb{P}X}(0) = \log M_{X-\mathbb{P}X}(0) = 0$.

The cumulants can be related to the moments by equating coefficients in power series expansions. Here is how it works for the first three cumulants.

$$\begin{aligned}
1 + \mu_2\lambda^2/2! + \mu_3\lambda^3/3! + O(\lambda^4) &= \exp\left(\sum_{k\in\mathbb{N}} \kappa_k\lambda^k/k!\right) \\
&= 1 + \left(\kappa_1\lambda + \kappa_2\lambda^2/2! + \kappa_3\lambda^3/3! + O(\lambda^4)\right) \\
&\quad + \left(\kappa_1\lambda + \kappa_2\lambda^2/2! + \kappa_3\lambda^3/3! + O(\lambda^4)\right)^2 \\
&\quad + \left(\kappa_1\lambda + \kappa_2\lambda^2/2! + \kappa_3\lambda^3/3! + O(\lambda^4)\right)^3 + O(\lambda^4) \\
&= 1 + \kappa_1\lambda + \kappa_2\lambda^2/2! + \kappa_3\lambda^3/3! + \left(\kappa_1^2\lambda^2 + \kappa_1\kappa_2\lambda^3\right) + \left(\kappa_1^3\lambda^3\right) + O(\lambda^4) \\
&= 1 + \lambda\left(\kappa_1\right) + \frac{\lambda^2}{2!}\left(\kappa_2 + 2\kappa_1^2\right) + \frac{\lambda^3}{3!}\left(\kappa_3 + 6\kappa_1\kappa_2\right) + O(\lambda^4).
\end{aligned}$$

It follows that $\kappa_1 = 0$ and $\kappa_2 = \mu_2$ and $\kappa_3 = \mu_3$.

Remark. Don't get too excited and leap to the conclusion that cumulants are the same as moments. If I hadn't centered the distribution to zero expected value then κ_1 would not be zero and μ_k would be a nasty-looking polynomial in $\kappa_1, \ldots, \kappa_k$. By repeated substitutions we could then write κ_k as another nasty-looking polynomial in the non-central moments $\mathbb{P}X, \ldots, \mathbb{P}X^k$. Even with the centering the remaining cumulants get messier: $\kappa_4 = X^2 - (\mathbb{P}X^2)^2$ and the expression for κ_{10} is a sum of 12 terms.

Back to tail probabilities. Remember that $\Lambda(x)$ is usually obtained by maximizing $\lambda x - L_{X-\mathbb{P}X}(\lambda)$ with respect to λ, with the task coming down to solving

$$t = L'_{X-\mathbb{P}X}(\lambda) = \kappa_2\lambda + \kappa_3\lambda^2/2! + \kappa_4\lambda^3/3! + \ldots \qquad .$$

This regularity suggests that the maximizing value λ_t be expressible as a power series $\sum_{k\in\mathbb{N}} a_k t^k/k!$. The a_k coefficients can be determined by another exercise in coefficient matching. First note that

$$\lambda_t = a_1 t + a_2\frac{t^2}{2!} + a_3\frac{t^3}{3!} + O(t^4), \quad \lambda_t^2 = a_1^2t^2 + a_1a_2t^3 + O(t^4), \quad \lambda_t^3 = a_1^3t^3 + O(t^4).$$

Thus

$$\begin{aligned} t &= \kappa_2\left(a_1t + a_2\frac{t^2}{2!} + a_3\frac{t^3}{3!}\right) + \frac{\kappa_3}{2!}\left(a_1^2t^2 + a_1a_2t^3\right) + \frac{\kappa_4}{3!}(a_1^3t^3) + O(t^4) \\ &= t\left(\kappa_2a_1\right) + \frac{t^2}{2!}\left(\kappa_2a_2 + \kappa_3a_1^2\right) + \frac{t^3}{3!}\left(\kappa_2a_3 + 3\kappa_3a_1a_2 + \kappa_4a_1^3\right) + O(t^4), \end{aligned}$$

implying

$$a_1 = 1/\kappa_2, \qquad a_2 = -\kappa_3/\kappa_2^3, \qquad a_3 = \text{something}.$$

It now follows that

$$\begin{aligned} \Lambda(t) &= t\left(a_1t + a_2t^2/2\right) - \frac{\kappa_2}{2!}\left(a_1^2t^2 + a_1a_2t^3\right) - \frac{\kappa_3}{3!}\left(a_1^3t^3\right) + O(t^4) \\ &= \frac{t^2}{2\kappa_2} - \frac{\kappa_3t^3}{6\kappa_2^3} + O(t^4) \qquad \text{near the origin.} \end{aligned}$$

The contributions from a_3 and κ_4 get absorbed into the $O(t^4)$.

Remark. You might be wondering why I bothered expanding λ_t as a cubic once I had realized the a_3 term would be absorbed into the $O(t^4)$. As a wise friend once advised me, it is always a good idea to expand an approximating series out to at least one term more than one thinks is necessary. A classic example of a failure to expand far enough can be found in the famous paper of Pearson (1900, page 165), which led to many years of argument between him and R. A. Fisher over the number of degrees of freedom for a χ^2 goodness of fit test. See Cochran (1952) for a very clear discussion of Pearson's error.

MGF::Binomial.local <5> **Example.** Suppose $X \sim \text{BIN}(n, p)$. Then

$$\mathbb{P}X = np, \quad \kappa_2 = \mathbb{P}(X - np)^2 = npq, \quad \kappa_3 = \mathbb{P}(X - np)^3 = npq(q - p)$$

so that

$$\Lambda(t) = \frac{t^2}{2npq} - \frac{t^3(q-p)}{6(npq)^2} + O(t^4) \qquad \text{near the origin.}$$

This approximation shows that $\Lambda(t) \geq t^2/(2npq)$ if $p \geq 1/2$ and $0 \leq t \approx 0$, in which case

$$\mathbb{P}\{X \geq np + t\} \leq e^{-\Lambda(t)} \leq \exp\left(-\frac{t^2}{2npq}\right) \qquad 0 \leq t \approx 0.$$

In fact, as will be shown in Section 3.7, the inequality holds for all $t \geq 0$ if $p \geq 1/2$. The local property implied by negative skewness suggests a subgaussian upper tail; the convexity of the ψ_{benn} function will transform the local suggestion into a global inequality. □

3.3 Normal

MGF::S:normal

The MGF method is cleanest for the normal distribution. As this Section will show, the method leads to bounds comparable to very sharp inequalities that can be derived using special properties of the normal.

MGF::normal <6> **Example.** If X has a $N(\mu, \sigma^2)$ distribution then $M(\lambda) = \exp(\lambda\mu + \sigma^2\lambda^2/2)$ is finite for all real λ. For $x \geq 0$ inequality <1> gives

$$\begin{aligned}\mathbb{P}\{X \geq \mu + \sigma x\} &\leq \inf_{\lambda \geq 0} \exp(-\lambda(\mu + \sigma x) + \lambda\mu + \lambda^2\sigma^2/2) \\ &= \exp(-x^2/2) \qquad \text{for all } x \geq 0,\end{aligned}$$

the minimum being achieved by $\lambda = x/\sigma$. Analogous arguments, with $X - \mu$ replaced by $\mu - X$, give an analogous bound for the lower tail,

$$\mathbb{P}\{X \leq \mu - \sigma x\} \leq \exp(-x^2/2) \qquad \text{for all } x \geq 0,$$

leading to the inequality $\mathbb{P}\{|X - \mu| \geq \sigma x\} \leq 2e^{-x^2/2}$, which shows that the distribution of X is concentrated near μ. □

Of course the algebra in the Example would have been a tad simpler if I had worked with the standardized variable $(X - \mu)/\sigma$. I did things the messier way in order make the point that if Y is any random variable (not necessarily normally distributed) for which there exist constants ν and τ (not necessarily the mean and variance) for which

$$M_Y(\lambda) = \mathbb{P}e^{\lambda Y} \leq e^{\nu\lambda + \lambda^2\tau^2/2} \qquad \text{for all } \lambda \geq 0$$

then

\E@ subg.upper.tail <7> $$\mathbb{P}\{Y \geq \nu + \tau x\} \leq e^{-x^2/2} \qquad \text{for } x \geq 0.$$

This inequality <7> is usually called a ***subgaussian bound for the upper tail***. However, as explained in Chapter 7, the term 'subgaussian' is also often used in several looser senses.

How good is the upper bound

\E@ normal.subg2 <8>
$$\bar{\Phi}(x) := \mathbb{P}\{Z \geq x\} \leq B(x) := e^{-x^2/2} \qquad \text{for } x \geq 0?$$

Clearly, not so good for x close to 0, because $\bar{\Phi}(0) = 1/2$ whereas $B(0) = 1$. As will be shown by Example <17>, it does capture the most important tail properties if x is very large, but how useful is it for intermediate values of x such as those that appear in typical statistical applications?

MGF::conf.int <9> **Example.** Anyone who has taken an introductory Statistics course knows that if $T \sim N(\theta, 1)$ under a $\mathbb{P}_\theta$ model then, to two decimal places accuracy,

$$\mathbb{P}_\theta\{T - 1.96 \leq \theta \leq T + 1.96\} = 2\bar{\Phi}(1.96) \approx 0.95.$$

That is, the range $T \pm 1.96$ is a 95% confidence interval for θ.

Consider the effect of using the upper bound $B(x)$ instead of $\bar{\Phi}(x)$ when constructing the confidence interval. It is certainly true that

$$\mathbb{P}_\theta\{T \pm c \text{ fails to contain } \theta\} \leq 2B(c) = 2\exp(-c^2/2),$$

so that the $\mathbb{P}_\theta$ probability that the interval $T \pm c$ fails to contain θ is smaller than $2B(c)$; replacement of $\bar{\Phi}$ by B leads to a "conservative confidence" assertion. Here are some examples (with values rounded to two decimal places):

x	1.64	1.96	2.45	2.58	2.72	3.26
$\bar{\Phi}(x)$	5%	2.5%	0.72%	0.5%	0.33%	0.06 %
$\exp(-x^2/2)$	25.85%	14.65%	5%	3.62%	2.5%	0.5 %

As judged by B, the range $T \pm 1.96$ will contain θ with probability at least 70.7%, which is not very comforting given that the nominal value is 95%. For the 90% interval, the conservative value, $1 - 2 \times 25.85\% = 48.3\%$, is even worse. It takes a great stretching of the imagination to describe either conclusion as 'not bad'.

There is another way to use the upper bound. Instead of stretching imagination we could stretch the interval, from $T \pm 1.96$ to $T \pm c$ with $c = 2.72$. For this interval, the B bound assures coverage of at least 95%. Looking on the bright side, I think the increase from 1.96 to 2.72 is not too high a price to pay for an appreciable relaxation of the modeling assumptions from normal to subgaussian.

The compromise looks even better when the failure probability is smaller. For example, under the strict $N(\theta, 1)$ assumption $T \pm 2.58$ is a 99% confidence interval for θ and $T \pm 3.26$ has, according to B, probability at least 99% of containing θ. □

Sharper tail bounds are possible if we exploit further properties of the normal density, $\phi(x) := (2\pi)^{-1/2}\exp(-x^2/2)$, properties that are not shared by all subgaussian distributions. In fact there is a literature going back over two centuries that contains numerous facts about $\bar{\Phi}$, including several upper and lower bounds. For example,

\E@ Laplace.0 <10> $$\left(x^{-1}-x^{-3}\right)\phi(x) < \bar{\Phi}(x) <^{-1} \phi(x) \qquad \text{for all } x>0,$$

which can be derived by integrating $\int_x^\infty$ through the pointwise bounds (for $r>0$)

$$-\frac{d}{dr}\left[(r^{-1}-r^{-3})\phi(r)\right] = (1-3r^{-4})\phi(r) < \phi(r)$$
$$< (1+r^{-2})\phi(r) = -\frac{d}{dr}\left[r^{-1}\phi(r)\right].$$

Inequality <10> is the first of a sequence of approximations essential due to Laplace (see Notes): for $k=0,2,4,\ldots$,

\E@ Laplace.k <11> $$p_{k+1}(1/x) < \mathcal{R}(x) := \bar{\Phi}(x)/\phi(x) < p_k(1/x) \qquad \text{for } x>0,$$

where p_k is a polynomial of degree $2k+1$: $p_0(r) := r$, $p_1(r) := p_0(r) - r^3$, $p_2(r) = p_1(r) + 3r^5$, $p_3(r) = p_2(r) - 15r^7$, ... See Problem [1] for the proof.

The function $\mathcal{R}(x) := \bar{\Phi}(x)/\phi(x)$ is often called the "Mills ratio", for not particularly compelling reasons (see the Notes). I find it more convenient to work with the reciprocal of $\mathcal{R}$,

\E@ rho.def <12> $$\rho(x) := 1/\mathcal{R}(x) = \phi(x)/\bar{\Phi}(x),$$

which is sometimes called the ***hazard rate*** for the $N(0,1)$ distribution because, for small positive δ,

$$\mathbb{P}\{x \le Z \le x+\delta \mid Z \ge x\} \approx \delta\rho(x) \qquad \text{if } Z \sim N(0,1).$$

Of course all the inequalities in <11> are rather useless for x near 0; the lower bounds $p_k(1/x)$ for odd k are all negative for $0 < x \le 1$. They are most useful for values of x much larger than 1, where they can be inverted to give upper and lower bounds for ρ. For example, the case $k=0$ gives $1-x^{-2} < x/\rho(x) < 1$, which inverts for small values of $y = 1/x$ to give

\E@ rho.0 <13> $$x < \rho(x) < x\left(1-y^2\right)^{-1} = x\left(1+y^2+y^4+\ldots\right) = x + x^{-1} + O(x^{-3}).$$

Similarly for $k=2$ we have $1-y^2+3y^4-15y^6 < x/\rho(x) < 1-y^2+3y^4$, which inverts to give

\E@ rho.2 <14> $$\rho(x) = x + x^{-1} - 2x^{-3} + O(x^{-5}) \qquad \text{as } x \to \infty.$$

We'll be needing both <13> and <14> in Section 3.4.

The literature also contains several bounds that do not blow up as $x \searrow 0$. For example,

\E@ more.bounds <15>
$$(3x + \sqrt{x^2+8})/4 < \rho(x) \le \left(x + \sqrt{x^2+4}\right)/2 \qquad \text{for all } x \in \mathbb{R},$$

the lower bound coming from Birnbaum (1942) and the upper bound from Sampford (1953). See Problems [4] and [3] for proofs.

Remark. The inequalities <15> are not accurate enough for moderate x to serve as a basis for numerical calculation of normal tail probabilities. These days, such results are of little interest once one accords ρ the status of a useful member of the standard repetoire of functions for which highly accurate approximations can be obtained in every decent mathematical or statistical computing package.

For $x > 1$ inequality <10> can be rewritten as

\E@ normal.tail2 <16>
$$\log \bar{\Phi}(x) = -x^2/2 - \log(x) - c_0 - \eta(x)$$

where $c_0 := \log\left(\sqrt{2\pi}\right)$ and $0 \le \eta(x) \le -\log(1 - x^{-2})$, which is $\le 2x^{-2}$ for $x \ge \sqrt{2}$. The bound $B(x)$ has captured the $-x^2/2$, which is much more important than the other terms when x becomes large, a fact illustrated by the next Example.

MGF::max.normal <17>
Example. Suppose $Z_1, \ldots, Z_n$ are random variables, each distributed $N(0,1)$ but, for the moment, not necessarily independent. Define $M_n := \max_{i \le n} Z_i$. A union bound gives some control for the tail:

$$\mathbb{P}\{M_n > x\} \le \sum\nolimits_i \mathbb{P}\{Z_i > x\} = n\bar{\Phi}(x) \le n \exp(-x^2/2) \qquad \text{for } x > 0.$$

In particular, $\mathbb{P}\{M_n > \sqrt{2\log(n) + 2r}\,\} \le e^{-r}$ for each $r \ge 0$. Roughly speaking, with high probability M_n should be not much bigger than $a_n := \sqrt{2\log n}$. This bound has the advantage of being unaffected by possible dependence between the Z_i's. It also has the disadvantage that it is sometimes excessively large. For example, in the extreme case where $Z_i = Z_1$ for all i any bound that involves n would be superfluous.

Remark. The union bound $\mathbb{P}\left(\cup_i A_i\right) \le \sum_i \mathbb{P} A_i$ is quite good if the events A_i are independent and $\sum_i \mathbb{P} A_i$ is small. See Problem [8]. It is part of the folklore that if the A_i's are 'almost' independent then the union bound is 'almost' quite good, with the meaning of 'almost' being problem specific.

If the Z_i's are actually independent the union bound can be refined:

$$\begin{aligned}
\mathbb{P}\{M_n \le x\} &= \prod\nolimits_i \mathbb{P}\{Z_i \le x\} = \left(1 - \bar{\Phi}(x)\right)^n \\
&= \exp\left(n \log(1 - \bar{\Phi}(x))\right) \\
&= \exp\left(-n\bar{\Phi}(x) - n \sum\nolimits_{j \ge 2} \left(\bar{\Phi}(x)\right)^j / j\right) \\
&= \exp\left(-n\bar{\Phi}(x) - R_n(x)\right) \qquad \text{where } 0 \le R_n(x) \le \frac{n\bar{\Phi}(x)^2}{2(1 - \bar{\Phi}(x))}.
\end{aligned}$$

\E@ indepN(0,1) <18>

If we choose x_n so that $n\bar{\Phi}(x_n)$ is suitably large then $\mathbb{P}\{M_n > x_n\}$ is small.

To get sharper bounds we need to take care of the $\log(x_n) + c_0$ contributions. It will turn out that most of the distribution of M_n concentrates in a very short interval centered at a point smaller than a_n. To that end consider x_n of the form $a_n - w_n/a_n$ with $|w_n|$ much smaller than a_n. Approximation <16> then gives

$$\begin{aligned}\log\left(n\bar{\Phi}(x_n)\right) &= \log n - (a_n^2 - 2w_n + w_n^2/a_n^2)/2 - \log a_n - c_0 \\ &\qquad - \log(1 - w_n/a_n^2) - \eta(x_n) \\ &= w_n - \log(a_n) - c_0 - \text{small remainder}\end{aligned}$$

where the remainder term is of order $O\left((1 + w_n^2)/a_n^2\right)$. Thus, for values $w_n = r + c_0 + \log a_n$ with, say, $|r| = O(\log a_n)$ we have

$$\mathbb{P}\{M_n \le a_n - (r + c_0 + \log a_n)/a_n\} = \exp\left(-B_n(r) - O(B_n(r)^2/n)\right).$$

where

$$B_n(r) := n\bar{\Phi}(x_n) = \exp\left(r + O(\epsilon_n^2)\right) \qquad \text{where } \epsilon_n := a_n^{-1}\log(a_n).$$

For example, when n is large enough, M_n has probability over 0.95 of lying in the range $a_n - (c_0 + \log a_n)/a_n \pm 4/a_n$. □

Remark. Inequalities <10> and <18> are the basis for the classical fact that $a_n(M_n - b_n)$, with $b_n = a_n - (\log\log n + \log(4\pi))/(2a_n)$, converges in distribution. See Leadbetter, Lindgren, and Rootzén (1983, Theorem 1.5.3).

In a digression from MGF methods, the next Section derives inequalities that are much more informative, for theoretical purposes, than <10> and <15>.

*3.4 Global behavior of the normal hazard rate

MGF::S:better-normal

Perhaps surprisingly, it pays to understand how ρ behaves over the whole real line, and not just for large x or for x near 0. For example, as a short discussion (at the end of the Section) of Stein's method for normal approximation will show, it is more convenient to deal with a single well-behaved function than having to argue differently for different ranges of x values. The proof of the following Theorem, which lists a few basic facts about ρ, also introduces some useful general tricks.

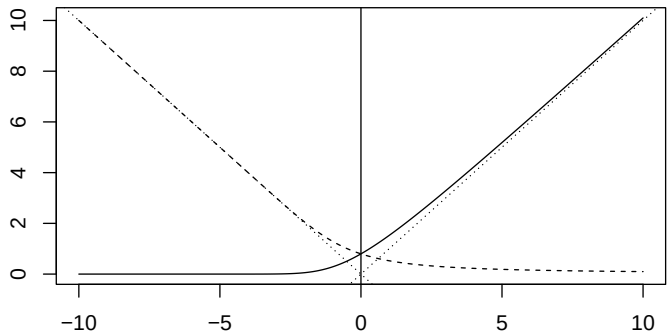
MGF::rho.facts <19> **Theorem.** *The function $\rho(x) := \phi(x)/\bar{\Phi}(x)$ is infinitely differentiable with:*

(i) The function $\log\rho$ is strictly concave on $\mathbb{R}$ and the function ρ is strictly {positive & increasing & convex} on $\mathbb{R}$, with $\rho(x) \to 0$ as $x \to -\infty$ and $\rho(x)/x \to 1$ as $x \to \infty$.

(ii) *The function* $r(x) := \rho(x) - x$ *is strictly* {*positive & decreasing & convex*} *on* $\mathbb{R}$ *with* $r(x) \to +\infty$ *as* $x \to -\infty$ *and* $r(x) = O(x^{-1})$ *as* $x \to \infty$.

□ *(iii)* $\dot{\rho}(x) = r(x)\rho(x) < 1$ *for all real* x *with* $r(0) = \rho(0) = \sqrt{2/\pi} \approx 0.798$.

The functions ρ (solid line), r (dashed line), and $\pm x$ (dotted line)



Proof. The trickiest part of the argument—establishing convexity of ρ—will be left to last. The rest is mostly calculus with just a few probability tricks.

First note that

$$1/\rho(x) = \int_0^\infty \phi(x+t)/\phi(x)\, dt$$

\E@ rr.rep <20>
$$= \int_0^\infty e^{-xt-t^2/2} dt = \sqrt{\pi/2}\, \mathbb{P} e^{-x|Z|} \qquad \text{where } Z \sim N(0,1).$$

Clearly $1/\rho$ must be strictly {decreasing & positive} and, by Section 2.3,

\E@ -log.rho <21>
$$\mathcal{L}(x) := -\log \rho(x) = x^2/2 + \log\sqrt{2\pi} + \log \bar{\Phi}(x)$$

must be strictly convex on the whole real line, ensuring that

\E@ ll.dot <22>
$$\dot{\mathcal{L}}(x) = x - \phi(x)/\bar{\Phi}(x) = x - \rho(x) = -r(x)$$

is strictly increasing and $\ddot{\mathcal{L}}(x) > 0$ for all real x. It follows that ρ must be strictly {increasing & positive} with $\lim_{x\to-\infty} \rho(x) = 0$ and $\lim_{x\to-\infty} \rho(x) = \infty$, and $\log \rho(x)$ strictly concave. Moreover, $\rho(x) - x = r(x)$ is strictly {decreasing & positive} on the whole real line.

Equality <21> also gives $-\dot{\rho}(x)/\rho(x) = \dot{\mathcal{L}}(x) = -r(x)$, that is,

\E@ rho.r <23>
$$\dot{\rho}(x) = \rho(x)r(x) = 1 - x^{-2} + O(x^{-4}) \qquad \text{as } x \to \infty,$$

the final assertion coming from equation <14>. Of course the limiting behavior does not establish the global property (iii) of the Theorem. A little probability trick comes to the rescue. Direct calculation shows that

$$\dot{\phi}(r) = -r\phi(r) \quad \text{AND} \quad \ddot{\phi}(r) = (r^2 - 1)\phi(r).$$

In integrated form: for all real x,

$$\int_x^\infty r\phi(r)\,dr/\bar\Phi(x) = \phi(x)/\bar\Phi(x) = \rho(x),$$
$$\int_x^\infty r^2\phi(r)\,dr/\bar\Phi(x) = \int_x^\infty \ddot\phi(r) + \phi(r)\,dr/\bar\Phi(x) = x\rho(x) + 1.$$

If P_x denotes the probability measure that has density $\phi(r)\{r \ge x\}/\bar\Phi(x)$ with respect to LEBESGUE measure on the real line then its expected value is $\rho(x)$ and its variance is $1 + x\rho(x) - \rho(x)^2 = 1 - r(x)\rho(x)$. As P_x is not degenerate its variance is nonzero, whence the asserted inequality for $r(x)\rho(x)$.

It remains to show that ρ is strictly convex. From <23> we get

$$\begin{aligned}\ddot\rho(x) &= r(x)\dot\rho(x) + \dot r(x)\rho(x)\\ &= r^2(x)\rho(x) + [\rho(x)r(x) - 1]\,\rho(x)\\ &= [\psi(x) - 1]\,\rho(x) \qquad \text{where } \psi(x) := r(x)\,[r(x) + \rho(x)]\,.\end{aligned}$$

\E@ mmdot.rho <24>

Notice that <23> implies $\psi(x) \to 1$ as $x \to \infty$. Differentiate again.

$$\begin{aligned}\dot\psi(x) &= \dot r(x)\,[r(x) + \rho(x)] + r(x)\,[\dot r(x) + \dot\rho(x)]\\ &= 2\dot r(x)r(x) + [\psi(x) - 1]\,\rho(x) \qquad \text{by <24>.}\end{aligned}$$

In the last line the $2\dot r(x)r(x)$ is strictly negative because it equals the derivative of the strictly decreasing function $r(x)^2$. Thus we have the strict inequality

$$\dot\psi(x) < [\psi(x) - 1]\,\rho(x) \qquad \text{for all } x \text{ in } \mathbb{R}.$$

\E@ psi.dot <25>

By <24>, to prove strict convexity for ρ it suffices to show $\psi(x) > 1$ for all x. To argue by contradiction (my least favorite method), suppose there were an x_0 for which $\psi(x_0) \le 1$. By <25>, we would have $\dot\psi(x_0) < 0$, implying existence of an x_1 slightly larger than x_0 at which $\psi(x_1) < 1$. As $\lim_{x\to\infty}\psi(x) = 1$, it would follow that the differentiable function ψ must achieve its infimum on $[x_1, \infty)$ at some x_2 where $\dot\psi(x_2) = 0$ but $\psi(x_2) \le \psi(x_1) < 1$, contradicting <25>. □

I claim that the results listed in Theorem <19> can simplify the discussion of other theory involving the normal distribution. The following discussion illustrates my point.

First a little background. One of the jewels of modern probability theory is the amazingly powerful method for obtaining normal approximations invented by Charles Stein (Stein, 1972, 1986). For a variety of random variables X his method provides bounds for $|\mathbb{P}h(X) - \gamma h|$ for a large family of bounded, measurable functions h on the real line. It works by means of a smoothing map κ defined by

$$\begin{aligned}H(r) &:= h(r) - \gamma h = \gamma^s\,[h(r) - h(s)]\,,\\ \phi(x)\kappa(x,h) &:= \gamma^r\{r < x\}H(r) = -\gamma^r\{r > x\}H(r).\end{aligned}$$

\E@ f.def <26>

The second representation comes from the fact that $\gamma H = 0$. In more traditional notation,

$$\phi(x)\kappa(x,h) := \int_{-\infty}^{x} H(r)\phi(r)\,dr = -\int_{x}^{\infty} H(r)\,\phi(r)\,dr.$$

Suppose $f(x) = \kappa(x,h)$. Typically h will be continuous at all but a finite number of points. At continuity points of h the function f is differentiable and $\dot\phi(x)f(x) + \phi(x)\dot f(x) = \phi(x)H(x)$. By virtue of the fact that $\dot\phi(x) = -x\phi(x)$, this equality simplifies to

\E@ Stein.de <27>

$$\dot f(x) - xf(x) = H(x) := h(x) - \gamma h \qquad \text{if } h \text{ is continuous at } x.$$

If X is a random variable then $\mathbb{P}\left(\dot f(X) - Xf(X)\right) = \mathbb{P}h(X) - \gamma h$. If we can show, using various calculus and probability tricks, that the expression on the left-hand side of this equality is small then we can deduce $\mathbb{P}h(X) \approx \gamma h$, that is, we will have a kind of normal approximation for X. In particular, if the left-hand side is zero for a large enough collection of h functions then X has a $N(0,1)$ distribution.

Stein's method depends on a small collection of facts that are summarized in the next Lemma. Initially, all we know is that $\|H\|_\infty := \sup_{x\in\mathbb{R}} |H(x)|$ is finite and that h is mostly continuous. Extra assumptions about h give extra smoothness properties for $\kappa(x,h)$.

You might find it helpful to keep two particular cases in mind while reading the proofs of (i) through (iv). For h_1 equal to the indicator function of an interval $(-\infty, x_0]$ we get

$$\phi(x)f_1(x) = \Phi(x \wedge x_0) - \Phi(x)\Phi(x_0).$$

The function f_1 is smooth except for a discontinuity in the first derivative at x_0. It can be used to prove the Berry-Esseen extension of the central limit theorem. See Ho and Chen (1978, §1) for a proof in the case of identically distributed summands, which they attributed to Stein himself. See also Hall and Barbour (1984) for related results with non-identically distributed summands.

For a smoother approximation to $(-\infty, x_0]$ we could use

$$h_2(r) = \{r \le x_0\} + \{x_0 < r < x_0 + \epsilon\}\,(1 + (x_0 - r)/\epsilon)$$

for some small, positive ϵ. This function is piecewise linear and LIPSCHITZ continuous, with $\|h_2\| = 1$ and $\|h_2\|_{Lip} = 1/\epsilon$. It appeared in a proof by Stein (1986, pp. 35–36). The corresponding $\kappa(x,h_2)$ is given by

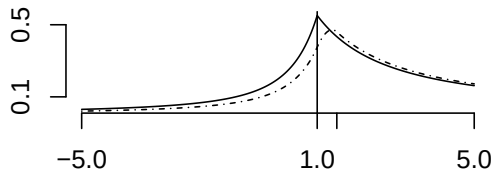
$$\phi(x)f_2(x) = \Phi(x \wedge x_1) - \Phi(x)\Phi(x_1) + \epsilon^{-1}\left[c_2\bar\Phi(x) - W(x)\right]$$

where

$$W(x) := \{x < x_1\}\Big(x_0\left[\Phi(x_1) - \Phi(x_0 \vee x)\right] + \phi(x_1) - \phi(x_0 \vee x)\Big)$$

and $c_2 := \lim_{x\to-\infty} W(x) = x_0\,[\Phi(x_1) - \Phi(x_0)] + \phi(x_1) - \phi(x_0)$.

Picture of f_1 (solid line), f_2 (dotdash line), for $x_0 = 1$ and $\epsilon = 0.5$:



Notice that f_1 is not differentiable at x_0, although it does have left- and right-derivatives at that point. The function f_2 is actually everywhere differentiable and it has a second derivative everywhere except at x_0 and $x_0 + \epsilon$.

MGF::Stein <28> **Lemma.** *(based on Stein, 1986, pp 25–28) Suppose $f(x) = \kappa(x, h)$ for a bounded measurable function h, as in <26>. Then*

MGF::xle0 *(i) For g defined by $g(r) := -h(-r)$ we have $f(x) = \kappa(-x, g)$. Thus it suffices to concentrate the analysis on a general $\kappa(x, h)$ for $x \geq 0$.*

MGF::bnd.f *(ii) $|f(x)| \leq \|H\|_\infty \min(1/\rho(x), 1/\rho(-x)) \leq \sqrt{\pi/2}\|H\|_\infty$ for each $x \in \mathbb{R}$. Consequently, $f(x) \to 0$ as $|x| \to \infty$.*

MGF::bnd.fdot *(iii) If h is continuous at a particular x then f is differentiable at that x with $|\dot f(x)| \leq 2\|H\|_\infty$.*

MGF::h.Lip *(iv) If h is* LIPSCHITZ *then so is $\dot f$ and $\|\dot f\|_{Lip} \leq 2\|h\|_{Lip}$. Consequently, $|f(x+z) - f(x) - z\dot f(x)| \leq z^2\|h\|_{Lip}$.* □

Proof. Assertion (iv) will require the most work; the other assertions are more straightforward.

Proof of (i). For $x = -y$ the replacement of h by the function $g(r) := -h(-r)$ and the symmetry of the $N(0,1)$ distribution give

$$\begin{aligned}\phi(-y)\kappa(-y,h) &= \gamma^r\gamma^s\{-r < -y\}\gamma^s\,[h(-r) - h(-s)]\\ &= -\gamma^r\gamma^s\{r > y\}\gamma^s\,[g(r) - g(s)] = \phi(y)\kappa(y,g).\end{aligned}$$

Thus $\kappa(-y,h) = \kappa(y,g)$. □

Proof of (ii). Define $C_0 := \|H\|_\infty$. Then <26> gives two upper bounds for $|f(x)|$:

$$\begin{aligned}|f(x)| &\leq \gamma^r\{r > x\}|H(r)|/\phi(x) \leq C_0\bar\Phi(x)/\phi(x) = C_0/\rho(x),\\ |f(x)| &\leq \gamma^r\{r < x\}|H(r)|/\phi(x) \leq C_0\Phi(x)/\phi(x) = C_0/\rho(-x).\end{aligned}$$

Thus $|f(x)| \leq C_0 \min(1/\rho(x), 1/\rho(-x)) \leq C_0/\rho(0)$, because ρ is an increasing function and $\max(x,-x) \geq 0$. If $|x| \to \infty$ then the smaller of $1/\rho(x)$ and $1/\rho(-x)$ tends to zero. □

Proof of (iii). As noted in <27>, if h is continuous at x then f is differentiable at x with $\dot f(x) = G(x) + H(x)$ where $G(x) := xf(x)$. Also, by (ii),

$$|\dot f(x)| \le |G(x)| + |H(x)| \le C_0 \min\left(|x|/\rho(x), |x|/\rho(-x)\right) + C_0.$$

For $x > 0$ use $x \le \rho(x)$; for $x < 0$ use $-x \le \rho(-x)$. (Or just appeal to (i) for $x < 0$.) □

Proof of (iv). Write C_1 for $\|h\|_{Lip} = \|H\|_{Lip}$.

The LIPSCHITZ continuity of h implies that it is also absolutely continuous, in the sense described by UGMTP §3.4, a property that implies existence of a measurable function ψ with $\sup_{x\in\mathbb{R}} |\psi(x)| \le C_1$ for which:

- h is differentiable with derivative $\psi(x)$ at $\mathfrak{m}$-almost all x.
- $h(b) - h(a) = \mathfrak{m}^t\{a < t < b\}\psi(t)$ for each bounded interval $[a, b]$.

For $x \in \mathbb{R}$ and $\delta > 0$ we have

$$|\dot f(x+\delta) - \dot f(x)| \le |G(x+\delta) - G(x)| + |H(x+\delta) - H(x)|$$

As G has a continous derivative $\dot G(x) = (1+x^2)f(x) + H(x)$ for which $G(x+\delta) - G(x) = \int_x^{x+\delta} \dot G(t)\, dt$ and $\|H\|_{Lip} = C_1$, it will suffice to show that $\sup_{x\in\mathbb{R}} |\dot G(x)| \le C_1$.

The main idea is to write both H and f as integrals involving ψ, derive a similar representation for $\dot G$, then use the bound on $|\psi|$ together with facts about ρ to bound $\dot G(x)$. Start with H:

$$\begin{aligned} H(r) &= \gamma^s \left(h(r) - h(s)\right) \\ &= \gamma^s \mathfrak{m}^t \psi(t) \left(\{s < t < r\} - \{r < t < s\}\right) \\ &= \mathfrak{m}^t \psi(t) \gamma^s \left(\{s < t\}\{t < r\} - \{r < t\}\{t < s\}\right) \qquad \text{by FUBINI} \\ &= \mathfrak{m}^t \psi(t) \left[\{t < r\}\Phi(t) - \{r < t\}\bar\Phi(t)\right]. \end{aligned}$$

\E@ H.rep <29>

Then for f, using the second representation in <26>:

$$\begin{aligned} -\phi(x) f(x) &= \gamma^r\{x < r\} H(r) \\ &= \mathfrak{m}^t \psi(t) \gamma^r \{x < r\} \left[\{t < r\}\Phi(t) - \{r < t\}\bar\Phi(t)\right] \left[\{t < x\} + \{x < t\}\right] \\ &= -\mathfrak{m}^t \psi(t) \left[\{t < x\}\Phi(t)\bar\Phi(x) + \{x < t\}\bar\Phi(t)\Phi(x)\right] \end{aligned}$$

\E@ f.rep <30>

because

$$\begin{aligned} \Phi(t)\gamma^r\{x < r\}\{t < r\}\{t < x\} &= \{t < x\}\Phi(t)\bar\Phi(x), \\ \{x < r\}\{r < t\}\{t < x\} &= 0, \\ \Phi(t)\gamma^r\{x < r\}\{t < r\}\{x < t\} &= \{x < t\}\Phi(t)\bar\Phi(t), \\ -\bar\Phi(t)\gamma^r\{x < r\}\{r < t\}\{x < t\} &= -\{x < t\}\bar\Phi(t)\left[\Phi(t) - \Phi(x)\right]. \end{aligned}$$

The terms involving $\Phi(t)\bar\Phi(t)$ cancel.

Combine <29> and <30> to get the representation

$$\begin{aligned}\dot{G}(x) = &\quad x\dot{f}(x) + f(x) = (1+x^2)f(x) + xH(x)\\ = &\quad -\mathfrak{m}^t\psi(t)\Phi(t)\{t<x\}\left[(1+x^2)\bar{\Phi}(x)/\phi(x) - x\right] +\\ &\quad -\mathfrak{m}^t\psi(t)\bar{\Phi}(t)\{x<t\}\left[(1+x^2)\Phi(x)/\phi(x) - x\right]\end{aligned}$$

The two functions of x are closely related:

$$\begin{aligned}\ell(x) &:= (1+x^2)\bar{\Phi}(x)/\phi(x) - x = \frac{1+x^2-x(x+r(x))}{\rho(x)} = \frac{1-xr(x)}{\rho(x)},\\ \ell(-x) &= (1+x^2)\Phi(x)/\phi(x) + x = \frac{1+xr(-x)}{\rho(-x)}.\end{aligned}$$

The function ℓ is strictly positive over the whole real line: trivially for $x \le 0$ because $r(x) > 0$ and by Theorem <19>(iii) because $1-xr(x) > 1-\rho(x)r(x) > 0$ for $x > 0$. Together with the elementary calculus facts that

$$\begin{aligned}\int_{-\infty}^{x} \Phi(t)\,dt &= \phi(x) + x\Phi(x) = \phi(x)\left[1 + x/\rho(-x)\right] = \phi(x)r(-x)/\rho(-x),\\ \int_{x}^{\infty} \bar{\Phi}(t)\,dt &= \phi(x) - x\bar{\Phi}(x) = \phi(x)\left[1 - x/\rho(x)\right] = \phi(x)r(x)/\rho(x),\end{aligned}$$

and the assumption that $|\psi(t)| \le C_1$, these equalities and the strict positivity of ℓ give

$$\begin{aligned}|\dot{G}(x)|/C_1 &\le \ell(x)\int_{-\infty}^{x}\Phi(t)\,dt + \ell(-x)\int_{x}^{\infty}\bar{\Phi}(t)\,dt\\ &= \frac{\left[1-xr(x)\right]\phi(x)r(-x) + \left[1+xr(-x)\right]\phi(x)r(x)}{\rho(x)\rho(-x)}\\ &= \frac{\phi(x)\left[\rho(-x) + x + \rho(x) - x\right]}{\rho(x)\rho(-x)} = \phi(x)\left(1/\rho(x) + 1/\rho(-x)\right)\\ &= \bar{\Phi}(x) + \Phi(x) = 1.\end{aligned}$$

Apparently there were several cancellations, due to the symmetry of the $N(0,1)$ distribution, hidden within the upper bound for $|\dot{G}(x)|$.

To bound $|f(x+z) - f(x) - z\dot{f}(x)|$ write it as an integral:

\E@ R(x,z).def <31>
$$|z\int_0^1 \dot{f}(x+tz) - \dot{f}(x)\,dt| \le |z|\int_0^1 \|\dot{f}\|_{Lip}\,t|z|\,dt.$$

□

3.5 Poisson

MGF::S:Poisson

The normal distribution represents the prototype for the class of subgaussian distributions. In a similar way the POISSON provides the prototype for a class of distributions that might be called "subPoisson". These distributions

behave like subgaussians for moderately large deviations from the mean but decrease only a little faster than the exponential further out in the tails. The BENNETT inequalities in Chapter 8 will provide further examples.

Recall that a random variable Y has a POISSON(θ) distribution if

$$\mathbb{P}\{Y = k\} = e^{-\theta}\theta^k/k! \qquad \text{for } k = 0, 1, \ldots.$$

The parameter θ must be strictly positive. The random variable $X = Y - \theta$ has a zero expected value with $\operatorname{var}(X) = \theta$ and

$$L_X(\lambda) = \theta(e^\lambda - 1 - \lambda) = \theta\mathbb{f}(\lambda) \qquad \text{for all } \lambda \in \mathbb{R}.$$

As explained in Section 3.1, we can derive both upper and lower tail bounds from the function

$$-\Lambda(y) = \inf_{\lambda\in\mathbb{R}} \left(L_X(\lambda) - y\lambda\right) = \theta \inf_{\lambda\in\mathbb{R}} \left(\mathbb{f}(\lambda) - \lambda y/\theta\right).$$

Notice the appearance of our friend $\mathbb{f}$ from Section 2.2. Its comrade $\mathbb{h}$ is coming soon.

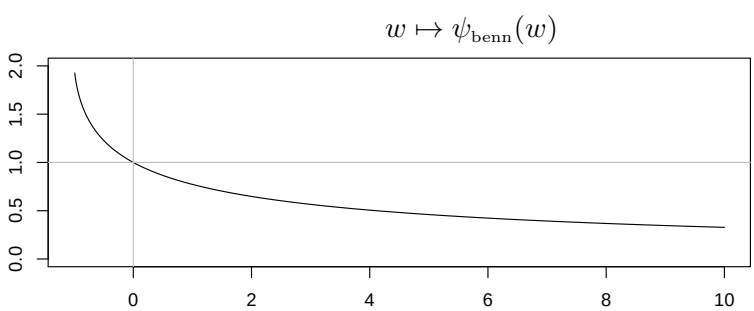
Temporarily write w for y/θ. Note that $\mathbb{f}(\lambda) - \lambda w = e^\lambda - 1 - \lambda(1+w)$ has derivative $e^\lambda - (1+w)$, which is zero at $\lambda = \log(1+w)$ if $w > -1$. If $w = -1$ the derivative is everywhere strictly positive, so that the infimum of -1 is approached as $\lambda \to -\infty$. If $w < -1$ then $\mathbb{f}(\lambda) - \lambda w = e^\lambda - 1 - \lambda(1+w)$, which approaches $-\infty$ as $\lambda \to -\infty$. In summary, $\inf_{\lambda\in\mathbb{R}} \left(\mathbb{f}(\lambda) - \lambda w\right)$ equals

\E@ Poisson.min <32>

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

Remark. If you have read Section 2.4 you will realize that I am here repeating the calculation that showed $\mathbb{h}$ is the Fenchel-Legendre conjugate of $\mathbb{f}$.

If you have read Section 2.2 you will also know that $\mathbb{h}(w) = \frac{1}{2}w^2\psi_{\text{benn}}(w)$ for $w \geq -1$, where $\psi_{\text{benn}}(\cdot)$ is a convex, decreasing function on $[-1, \infty)$ with $\psi_{\text{benn}}(0) = 1$. For large w the value of $\psi_{\text{benn}}(w)$ decreases like $2w^{-1}\log(w)$.



Thus

$$\Lambda(y) = \theta\mathbb{h}(y/\theta) = \begin{cases} \dfrac{y^2}{2\theta}\psi_{\text{benn}}(y/\theta) & \text{if } y \geq -\theta \\ \infty & \text{if } y < -\theta \end{cases},$$

which translates into

$$\begin{aligned}
\mathbb{P}\{X \ge x\} &\le \exp\left(-\frac{x^2}{2\theta}\psi_{\text{benn}}(x/\theta)\right) \qquad \text{for } x \ge 0 \\
\mathbb{P}\{X \le -x\} &\le \exp\left(-\frac{x^2}{2\theta}\psi_{\text{benn}}(-x/\theta)\right) \qquad \text{for } 0 \le x \le \theta \\
\mathbb{P}\{X \le -x\} &\le 0 \qquad \text{for } x > \theta.
\end{aligned}$$

The third inequality is reassuring because $\mathbb{P}\{X < -\theta\} = 0$. The first inequality shows that the upper tail decreases like a subgaussian in the range $0 \le x \ll \theta$, because $\psi_{\text{benn}}(x/\theta) \approx 1$ for x/θ near 0, but that the tail decay becomes more like $\exp(-x\log(x/\theta))$ further out into the tail. The inequality for the lower tail is more interesting, because $\psi_{\text{benn}}(w) > 1$ for $-1 \le w < 0$. The lower tails drop off even faster than one might expect from the $N(\theta,\theta)$ approximation to the POISSON(θ). This can be interpreted as a skewness effect: $\mathbb{P}X^3$ is the coefficient of $\lambda^3/3!$ in the power series expansion of

$$\begin{aligned}
\mathbb{P}e^{\lambda X} &= \exp\Big(\theta(e^\lambda - 1 - \lambda)\Big) \\
&= 1 + \theta\left(\frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots\right) + \frac{\theta^2}{2!}\left(\frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots\right)^2 + \dots,
\end{aligned}$$

which is positive. The distribution of $X - \theta$ puts more mass to the right of the origin than the $N(0,\theta)$. That fact slows down the decay in the upper tails but improves the rate of decay in the lower tail.

The MGF tail bound for the POISSON does not quite capture the actual behavior of the probabilities. The deficiency parallels what happens with the normal, where $\bar{\Phi}(x)$ decreases like $e^{-x^2}/\left(\sqrt{2\pi}x\right)$ for large x but the MGF method captures only the $\exp(-x^2/2)$. I hope you drew the conclusion from Section 3.3 that the failure was not fatal.

For $X = Y - \theta$ with $Y \sim$ POISSON(θ), the $\exp(-\theta\hbar(y/\theta))$ tail bound compares favorably with the probability calculated by means of the STIRLING formula (see Section 2.5) for $k \in \mathbb{N}$:

$$k! = \sqrt{2\pi}k^{k+1/2}e^{-k+r_k} \qquad \text{where } \frac{1}{12k+1} < r_k < \frac{1}{12k}.$$

If $k = \theta + y$ then

$$\begin{aligned}
\log\left(\sqrt{2\pi k}\,\mathbb{P}\{Y = k\}\right) &= -\theta + k\log(\theta) - k\log(k) + k - r_k \\
&= y - (\theta + y)\log(1 + y/\theta) - r_k \\
&= -\theta\hbar(y/\theta) - r_k.
\end{aligned}$$

Once again the MGF method has successfully captured the most important term, $-\theta\hbar(y/\theta)$, in the exponent.

3.6 Gamma and chi-squared

MGF::S:Gamma

Suppose X has a GAMMA(α) distribution, the probability measure on $\mathbb{R}^+$ that has density $f_\alpha(x) = x^{\alpha-1}e^{-x}/\Gamma(\alpha)$ with respect to LEBESGUE measure. The positive parameter α is often called the *shape parameter*. The expected value and variance of X both equal α and

$$M_X(\lambda) = \int_0^\infty \frac{x^{\alpha-1}e^{-x(1-\lambda)}}{\Gamma(\alpha)}\,dx = (1-\lambda)^{-\alpha}\{\lambda < 1\} + \infty\{\lambda \geq 1\}.$$

Thus $L_{X-\alpha}(\lambda) = -\alpha\lambda - \alpha\log(1-\lambda)$ for $\lambda < 1$, so that

\E@ upper.gamma.tail <33>
$$\mathbb{P}\{X \geq \alpha + t\} \leq \exp\Big(-\Lambda(t)\Big) \qquad \text{for } t \geq 0$$

\E@ lower.gamma.tail <34>
$$\mathbb{P}\{X \leq \alpha - t\} \leq \exp\Big(-\Lambda(-t)\Big) \qquad \text{for } \alpha > t \geq 0$$

where, for $\alpha + y > 0$,

$$\Lambda(y) = \sup_{\lambda<1} \lambda(\alpha+y) + \alpha\log(1-\lambda) = y - \alpha\log(1+y/\alpha).$$

The maximum is achieved at $\lambda = y/(\alpha+y)$. For $t \geq 0$ this gives

$$\log\mathbb{P}\{X \geq \alpha+t\} = \log\mathbb{P}\{X \geq t\} \leq -\Lambda(t) \approx \begin{cases} -t^2/(2\alpha) & \text{if } t \text{ is near } 0 \\ -t & \text{if } t \text{ is large} \end{cases}$$

and for $0 \leq t < \alpha$ it gives

$$\log\mathbb{P}\{X \leq \alpha - t\} = \log\mathbb{P}\{X \leq -t\} \leq -\Lambda(-t) \leq -t^2/(2\alpha).$$

The lower tail is actually subgaussian.

Boucheron, Lugosi, and Massart (2013, page 28) pointed out that the tail can also be bounded by first using an upper bound for the logMGF of $X-\alpha$: if $0 \leq \lambda < 1$ then

\E@ gamma.BML <35>
$$\alpha^{-1}\log\mathbb{P}e^{\lambda(X-\alpha)} = -\log(1-\lambda) - \lambda = \sum\nolimits_{i\geq 2}\frac{\lambda^i}{i} \leq \sum\nolimits_{i\geq 2}\frac{\lambda^i}{2} = \frac{\lambda^2}{2(1-\lambda)}.$$

They referred to this inequality as a $\Gamma_+(\alpha, 1)$ bound ('subgamma on the right tail'). As I'll only work with this tail I'll drop the 'on the right' qualifier. They also introduced a second parameter, for scaling.

MGF::subGamma.def <36>
Definition. *For constants $\alpha > 0$ and $\beta > 0$, interpret $W \in$ SUBGAMMA(α, β) to mean*

$$\mathbb{P}e^{\lambda W} \leq \exp\left(\frac{\alpha\lambda^2/2}{1-\beta\lambda}\right) \qquad \text{for } 0 \leq \lambda < 1/\beta.$$

□ *Abbreviate* SUBGAMMA$(\alpha, 1)$ *to* SUBGAMMA(α).

Remarks.

(i) If $W \in \text{subGamma}(\alpha)$ then

$$M_W(\lambda) = 1 + \lambda \mathbb{P}W + o(\lambda) \le \exp\left(\alpha\lambda^2 + o(\lambda^3)\right) = 1 + O(\lambda^2)$$

for small, positive λ, which implies $\mathbb{P}W \le 0$.

(ii) Notice that $W \in \text{subGamma}(\alpha, \beta)$ iff $W/\beta \in \text{subGamma}(\alpha/\beta^2)$. To me it seems cleaner to derive general theory with $\beta = 1$ and then deduce the corresponding $\text{subGamma}(\alpha, \beta)$ facts by rescaling. For example, if X is χ^2_k distributed then $X/2 \sim \text{Gamma}(k/2)$, which implies $X/2 - k/2 \in \text{subGamma}(k/2)$, and hence $X - k \in \text{subGamma}(2k, 2)$.

MGF::subGamma.tail <37>

Theorem. *If $W \in \text{subGamma}(\alpha)$ and $t \ge 0$ then*

$$\mathbb{P}\{W \ge t\} \le e^{-H_1(t,\alpha)} \le e^{-H_2(t,\alpha)}$$

where

$$H_1(t,\alpha) := \left(t + \alpha - \sqrt{2t\alpha + \alpha^2}\right) = \frac{t^2}{(t+\alpha) + \sqrt{2t\alpha + \alpha^2}} \quad ,$$

$$H_2(t,\alpha) := \frac{t^2}{2(t+\alpha)} \quad .$$

Remark. For future reference, here is the rescaled version of the Theorem: if $W \in \text{subGamma}(\alpha, \beta)$ and $t \ge 0$ then

$$\mathbb{P}\{W \ge t\} \le e^{-H_1(t/\beta, \alpha/\beta^2)} = \exp\left(\frac{-t^2}{\alpha + \beta t + \sqrt{\alpha^2 + 2t\alpha\beta}}\right)$$
$$\le e^{-H_2(t/\beta, \alpha/\beta^2)} = \exp\left(\frac{-t^2}{2(\alpha + \beta t)}\right).$$

Proof.

$$\begin{aligned}\log \mathbb{P}\{W \ge t\} &= \inf_{0 \le \lambda < 1} \left(-t\lambda + \frac{\alpha\lambda^2}{2(1-\lambda)}\right) \\ &= \inf_{0 < s \le 1} \left(-t(1-s) + \alpha\frac{1 - 2s + s^2}{2s}\right) \\ &= \inf_{0 < s \le 1} \left((t + \alpha/2)s - (t+\alpha) + \alpha/(2s)\right)\end{aligned}$$

The replacement of λ by $1 - s$ makes it easier to calculate the derivative, $t + \alpha/2 - \alpha/(2s^2)$, which is zero when $s = \sqrt{\alpha/(2t+\alpha)}$. That value gives the first expression for H_1. The final inequality comes from $\alpha^2 + 2\alpha t \le (\alpha + t)^2$. □

MGF::weighted.chi2 <38>

Example. (Laurent and Massart, 2000, Lemma 1) Consider the weighted sum $W = \sum_{j=1}^k a_j(Z_j^2 - 1)$ where $a = (a_1, \ldots, a_k) \in \mathbb{R}^k_+$ and $Z = (Z_1, \ldots, Z_k)$ has a $N(0, I_k)$ distribution. As usual, define

$$|a|_\infty := \max_j |a_j| \quad \text{AND} \quad |a|_2 := \sqrt{\sum\nolimits_j a_j^2}\,.$$

Each Z_j^2 has a χ_1^2 distribution which gives

$$a_j(Z_j^2 - 1) \in \text{subGamma}(2a_j^2, 2a_j) \subset \text{subGamma}(2a_j^2, 2|a|_\infty)$$

Consequently $W \in \text{subGamma}(2|a|_2^2, 2|a|_\infty)$ and, for example,

$$\mathbb{P}\{W \geq t\} \leq \exp\left(\frac{-t^2}{4|a|_2^2 + 4|a|_\infty t}\right) \qquad \text{for } t \geq 0.$$

Laurent and Massart rearranged the inequality into a form like <39> (see below) and also derived a companion bound for the lower tail. □

To compare the tail bounds from <33> and Theorem <37> it helps to isolate the effect of α by writing $\Lambda(t) = \alpha R_0(t/\alpha)$ and $H_j(t, \alpha) = \alpha R_j(t/\alpha)$ for $j = 1, 2$, for $x \geq 0$. Then

$$\begin{aligned} 0 \leq R_0(t) &:= t - \log(1+t) \\ &\leq R_1(t) := 1 + t - \sqrt{1+2t} = \frac{t^2}{1+t+\sqrt{1+2t}} \\ &\leq R_2(t) := t^2/(2t+2). \end{aligned}$$

The inequalities are all strict for $t > 0$. All three functions R_0, R_1, and R_2 have first derivatives that are positive and increasing; all are convex and strictly increasing. Near the origin $R_0(t) = t^2/2 - t^3/3 + 0(t^3)$ and both $R_1(t)$ and $R_2(t)$ behave like $t^2/2 - t^3/2 + o(t^3)$. As $t \to \infty$ both $R_0(t)/t$ and $R_1(t)/t$ converge to 1 but $R_2(t) \to 1/2$. If one is not too worried about the constants in the exponent there is not much difference between the three tail bounds.

It might appear that there is little point in recording the H_1 tail bound when it differs so little from the H_2 tail bound. However H_1 does give a more pleasing result if we rearrange the bound by solving $R_1(t) = w$ for a fixed $w > 0$ to get $t = w + \sqrt{2w}$, the larger of the two roots of the quadratic $(1+t-w)^2 = 1+2t$. Consequently, $\alpha R_1(t/\alpha) = w$ if $t = w + \sqrt{2\alpha w}$. With such a change of variable, the H_1 form of inequality from Theorem <37> takes the neat form (Boucheron et al., 2013, page 29)

\E@ BLM29 <39> $$\mathbb{P}\{W \geq w + \sqrt{2\alpha w}\,\} \leq e^{-w} \qquad \text{for } w \geq 0 \text{ if } W \in \text{subGamma}(\alpha).$$

Remark. You should carry out the analogous calculation for R_2. The result is not as elegant or useful.

In particular, if $X \sim \text{GAMMA}(\alpha)$ then

\E@ root.gamma <40>
$$\begin{aligned} \mathbb{P}\{\sqrt{X} \geq \sqrt{\alpha} + \sqrt{w}\} &= \mathbb{P}\{X \geq \alpha + 2\sqrt{\alpha w} + w\} \\ &\leq \mathbb{P}\{X - \alpha \geq w + \sqrt{2\alpha w}\,\} \leq e^{-w} \qquad \text{for } w \geq 0. \end{aligned}$$

Substituting t for $\sqrt{w}$ we get $\mathbb{P}\{\sqrt{X} \geq \sqrt{\alpha} + t\} \leq e^{-t^2}$, an example of a "subgaussian bound" for an upper tail beyond a point strictly larger than the mean: the JENSEN inequality gives $\mathbb{P}\sqrt{Y} < \sqrt{\alpha}$.

MGF::chi2 <41> **Example.** If $W \sim \chi^2_k$ then $W/2 \sim \text{GAMMA}(k/2)$ and inequality <40> implies

$$\mathbb{P}\{\sqrt{W} \geq \sqrt{k} + t\} \leq e^{-t^2/2} \qquad \text{for } t \geq 0.$$

In particular, $\mathbb{P}\{\sqrt{W} \geq 2\sqrt{k}\,\} \leq e^{-k/2}$ is a convenient bound when precise constant don't matter and k is large.

As an application, suppose $Z_1, \ldots, Z_n$ are independent $N(0, I_d)$ random vectors. Then we have $\mathbb{P}\,\|Z_i\| \leq \sqrt{\mathbb{P}\,\|Z_i\|^2} = \sqrt{d}$ and $\sum_{i\leq n} \|Z_i\|^2 \sim \chi^2_{nd}$. From the inequality $n^{-1}\sum_{i\leq n}\|Z_i\| \leq \sqrt{\sum_{i\leq n}\|Z_i\|^2/n}$ it follows that

$$\mathbb{P}\{n^{-1}\sum\nolimits_{i\leq n}\|Z_i\| \geq 2\sqrt{d}\,\} \leq e^{-nd/2},$$

a neat little bound that is useful in high-dimensional statistical theory. See Wu and Zhou (2019, Section 9), for example. □

3.7 Binomial

MGF::S:Binomial

The BINOMIAL distribution behaves a little like the POISSON. It is also a prototype for other inequalities involving sums of bounded random variables.

Remember that X has a $\text{BIN}(n, p)$ distribution if

$$\mathbb{P}\{X = k\} = \binom{n}{k} p^k q^{n-k} \qquad \text{for } k = 0, 1, \ldots, n.$$

Here and subsequently I write q for $1 - p$. The distribution has expected value np, variance npq, and $M_X(\lambda) = (q + pe^\lambda)^n$. The random variable $n - X$ has a $\text{BIN}(n, q)$ distribution. Thus

\E@ upper.lower <42> $$\mathbb{P}\{X \leq np - t\} = \mathbb{P}\{n - X \geq nq + t\}.$$

That is, the lower tail for the $\text{BIN}(n, p)$ corresponds exactly to the upper tail for the $\text{BIN}(n, q)$.

Here is the main result: If $X \sim \text{BIN}(n, p)$ then

\E@ Bin.upper <43>
$$\begin{aligned}\mathbb{P}\{X \geq np + t\} &\leq \exp\left(-np\hbar(t/np) - nq\hbar(-t/nq)\right) \\ &= \exp\left(-\frac{t^2}{2npq} g_p(x)\right) \qquad \text{for } 0 \leq t \leq nq \\ &\qquad \text{where } g_p(t) := q\psi_{\text{benn}}\left(\frac{t}{np}\right) + p\psi_{\text{benn}}\left(\frac{-t}{nq}\right).\end{aligned}$$

From equality <42> the companion inequality for the lower tail is

$$\mathbb{P}\{X \leq np - t\} \leq \exp\left(-\frac{t^2}{2npq} g_q(t)\right) \qquad \text{for } 0 \leq t \leq np.$$

It is merely a matter of swapping the roles of p and q.

Remarks.

(i) See Section 2.2 if you have forgotten about ψ_{benn}.

(ii) Note that

$$g_p(t)/(2npq) \to \psi_{\text{benn}}(t/\theta)/(2\theta) \qquad \text{if } n \to \infty \text{ and } np \to \theta \in \mathbb{R}^+.$$

Not surprisingly we then recover the MGF tail bounds for the POISSON(θ) distribution.

(iii) This $g_p(t)$ is the same function as the $g(t, n, p)$ in Section 2.5, which derived sharp approximations for $\mathbb{P}\{X = k\}$ by means of the STIRLING approximation: for $k = np + t$,

$$\mathbb{P}\{X = k\} = \frac{\exp\left[-t^2 g_p(t)/(2npq) + O\left(k^{-1} + (n-k)^{-1}\right)\right]}{\sqrt{2\pi n(p + t/n)(q - t/n)}}.$$

As happened with the $N(0, 1)$, the MGF method captures the main term in the exponent but misses the square root term in the denominator.

(iv) The bound $t \le nq$ is not really necessary, because $\mathbb{P}\{0 \le X \le n\} = 1$. It merely serves to ensure that $t/(np)$ and $-t/(nq)$ are both ≥ -1, so that we don't have to worry about $\psi_{\text{benn}}()$ taking the value $+\infty$. We could also let the the definition of $L_{X-np}(t)$ take care of the difficulty by having it take the value $+\infty$ when $t < -np$ or $t > nq$. Compare with the calculation for the POISSON in Section 3.5.

(v) The TAYLOR expansion $\hbar(x) = x^2/2! - x^3/3! + O(x^4)$ gives

$$np\hbar(t/np) + nq\hbar(-t/nq) = \frac{t^2}{2npq} - \frac{t^3(q-p)}{6(npq)^2} + O(t^4) \qquad \text{for } t \text{ near } 0,$$

which agrees with the calculations in Section 3.2.

(vi) As explained in Section 2.5, the convexity of ψ_{benn} gives the inequality

$$g_p(t) \ge \psi_{\text{benn}}\left(\frac{qt}{np} - \frac{pt}{nq}\right) = \psi_{\text{benn}}\left(\frac{t(q-p)}{npq}\right) \ge 1 \qquad \text{if } p \ge 1/2.$$

Thus, if $p \ge 1/2$, the upper tail is less than $\exp\left(-t^2/(2npq)\right)$, a clean subgaussian bound with scale parameter $\sqrt{npq}$. (As commented in Section 2.5, this subgaussian fact can also be interpreted as a skewness effect.) If $p < 1/2$ the upper tail is still subgaussian (because $\psi_{\text{benn}}(-t/(nq)) \ge 1$) but with a larger scale parameter $\sqrt{nq}$.

Proof (of inequality <43>). For Λ with $t \ge 0$ we need to find the supremum over $\mathbb{R}^+$ of

$$\mathcal{L}(\lambda) := (t + np)\lambda - n\log(q + pe^\lambda),$$

which has derivative

$$\dot{\mathcal{L}}(\lambda) = (t + np) - npe^\lambda/(q + pe^\lambda).$$

If $t = nq$ then $\dot{\mathcal{L}}(\lambda) > 0$ on $\mathbb{R}^+$ and $\mathcal{L}(\lambda) = n\log\left(e^\lambda/(q + pe^\lambda)\right)$, so that the supremum $n\log(1/p)$ is approached as $\lambda \to \infty$. The final bound then reduces to $\mathbb{P}\{X \ge n\} \le p^n$, which is actually true with equality.

If $0 \le t < nq$ then the maximum is achieved at the λ for which $\dot{\mathcal{L}}(\lambda) = 0$, that is, when $(t+np)(q+pe^\lambda) = npe^\lambda$. The algebra is then simplified a trifle if we write z_1 for $t/(np)$ and z_2 for $t/(nq)$. The equation becomes

$$(1+z_1)q = e^\lambda\,[1-p(1+z_1)] = e^\lambda q(1-z_2)$$

because $p(1+z_1)+q(1-z_2) = 1$. That is, the maximizing λ is given by $e^\lambda = (1+z_1)/(1-z_2)$ and

$$\begin{aligned}\Lambda(t) &= np(1+z_1)\log\left(\frac{1+z_1}{1-z_2}\right) - n\log\left(q+p(1-z_1)/(1-z_2)\right)\\ &= np(1+z_1)\log(1+z_1) - np(1+z_1)\log(1-z_2)\\ &\quad - n\log\left(q(1-z_2)+p(1+z_1)\right) + n\log(1-z_2)\\ &= np\left(\mathbb{h}(z_1)+z_1\right) - \log(1) + n\left(1-p(1+z_1)\right)\log(1-z_2)\\ &= np\left(\mathbb{h}(z_1)+t/(np)\right) + nq\left(\mathbb{h}(-z_2)-t/(nq)\right),\end{aligned}$$

□ which simplifies to the $np\mathbb{h}(z_1) + nq\mathbb{h}(-z_2)$ for the first line of <43>.

Now let me move a little beyond the BINOMIAL to show that there are several other distributions, with the same expected value as the $\text{BIN}(n,p)$, that share the tail bounds for the BINOMIAL. In this Chapter these bounds are derived by means of pointwise inequalities for MGFs. As such they leave open the question of whether analogous inequalities would also hold for the exact tail probabilities, not just their MGF-derived upper bounds. Chapter 4 will return to this question

MGF::PoisBin <44> **Example.** Suppose $S \sim \text{PBIN}(p_1,\ldots,p_n)$, that is, S is a sum of independent random variables $Y_1+\cdots+Y_n$, with $Y_i \sim \text{BER}(p_i)$ for possibly different p_i's. Define $\overline{p} = n^{-1}\sum_i p_i$. Then

$$M_S(\lambda) = \prod\nolimits_{i\le n}\left(q_i + p_i e^\lambda\right) = \exp\left(\sum\nolimits_{i\le n}\log(q_i+p_ie^\lambda)\right).$$

Concavity of the log() function shows that

\E@ PB.MGF <45> $$n^{-1}\sum\nolimits_{i\le n}\log(q_i+p_ie^\lambda) \le \log\left(n^{-1}\sum\nolimits_{i\le n}(q_i+p_ie^\theta)\right) = \log\left(\overline{q}+\overline{p}e^\theta\right).$$

Thus $M_S(\theta) \le M_W(\theta)$ where $W \sim \text{BIN}(n,\overline{p})$ and $\overline{q} = 1-\overline{p}$. It follows that

$$\mathbb{P}\{S \ge n\overline{p}+x\} \le \exp\left(-\frac{x^2}{2n\overline{p}\overline{q}}g_{\overline{p}}(x)\right) \qquad \text{for } 0 \le x \le n\overline{q},$$

□ with a similar bound for the lower tail.

The convexity idea from the previous Example can be pushed even further.

MGF::Hoeffding <46> **Example.** Suppose $T = Y_1+\cdots+Y_n$, a sum of independent random variables Y_i with $0 \le Y_i \le 1$ and $\mathbb{P}Y_i = p_i$ for each i, and $n\overline{p} = \sum_{i=1}^n p_i$. By convexity of the exp() function,

$$e^{\lambda Y_i} \le (1-Y_i) + Y_i e^\lambda \qquad \text{for each real } \lambda.$$

The inequality holds for all possible realizations of Y_i. Equality is achieved at $Y_i \in \{0,1\}$. In particular, equality holds when $Y_i \sim \text{BER}(p_i)$, as in the previous Example. Take expectations.

$$\mathbb{P}e^{\lambda Y_i} \le (1-p_i) + p_i e^{\lambda} = q_i + p_i e^{\lambda} \qquad \text{for each real } \lambda.$$

By independence,

$$M_T(\lambda) = \prod\nolimits_{i\le n} \mathbb{P}e^{\lambda Y_i} \le \prod\nolimits_{i\le n} \left(q_i + p_i e^{\lambda}\right) = M_S(\lambda) \le M_W(\lambda),$$

with $S \sim \text{PBIN}(p_1,\ldots,p_n)$ and $W \sim \text{BIN}(n,\overline{p})$, as in the previous Example. Thus

$$\mathbb{P}\{T \ge n\overline{p} + x\} \le \exp\left(-\frac{x^2}{2n\overline{p}\overline{q}} g_{\overline{p}}(x)\right) \qquad \text{for } 0 \le x \le n\overline{q},$$

□ a result due to Hoeffding (1963, Theorem 1).

3.8 Sampling and the hypergeometric

MGF::S:hypergeometric

Both Example <44> and Example <46> involved sums of independent random variables. The MGF approach can also work when there is dependence between the summands, although the argument becomes a little more delicate.

MGF::hyper <47> **Example.** Suppose $U = \{u_1,\ldots,u_N\}$ is a finite set, an urn if you like to think that way. In that interpretation the u_i's are the balls. Suppose exactly R of the balls are colored red and the other $B = N - R$ are colored black. If n balls are sampled without replacement then each subset of U with size n has probability $1/\binom{N}{n}$ of being selected and the number of red balls T_n in the sample has a hypergeometric distribution, $\text{HYPER}(n,R,B)$, meaning that

$$\mathbb{P}\{T_n = k\} = \binom{R}{k}\binom{B}{n-k}\Big/\binom{N}{n}$$

for each nonnegative integer k such that $k \le R$ and $n - k \le B$.

If the sampling is carried out with replacement then the number of red balls in the sample, S_n, has a $\text{BIN}(n,p)$ distribution, where $p = R/N$.

Elementary calculations (Pitman, 1993, Section 3.6) show that

$$\begin{aligned} \mathbb{P}T_n &= \mathbb{P}S_n = np \\ \text{var}(T_n) &= np(1-p)(N-n)/(N-1) < \text{var}(S_n) = np(1-p). \end{aligned}$$

If n is much smaller than N then there is actually not much difference between $\text{HYPER}(n,R,B)$ and $\text{BIN}(n,p)$: if a ball is selected then returned to the urn, it is unlikely to be selected again if n/N is very small. If n/N is not so small then, judging by the variances, $\text{HYPER}(n,R,B)$ is more concentrated around np than $\text{BIN}(n,p)$. A beautiful result by Hoeffding (1963,

Section 6) adds some precision to this intuition. He showed that for each convex function f on the real line,

\E@ Urn.Jensen <48>

$$\mathbb{P}f(T_n) \le \mathbb{P}f(S_n).$$

In particular, the choice $f(x) = e^{\lambda x}$ shows that $M_{T_n}(\lambda) \le M_{S_n}(\lambda)$ for all real λ. Any tail bound for the hypergeometric obtained via the MGF argument must therefore be smaller than the corresponding MGF tail bound for the BINOMIAL.

To be more precise, Hoeffding's result didn't involve red balls and black balls. It worked for every function $g : U \to \mathbb{R}$. (The special case where $g(u_i) = 1$ for a red ball an $g(u_i) = 0$ for a black ball get us back to the hypergeometric.) That is, we can take $X_1, \ldots, X_n$ to be a sample from U without replacement and $Y_1, \ldots, Y_n$ to be a sample with replacement. If we define $T_n := \sum_{i \le n} g(X_i)$ and $S_n := \sum_{i \le n} g(Y_i)$ then inequality <48> will still hold for every convex f.

I had some trouble digesting Hoeffding's proof. Even after working through the details I could not have explained to anyone why the method worked. Subsequently I stumbled on a proof by Le Cam (1986a, page 534), which involved a much more intuitive explanation, reducing everything to the JENSEN inequality. Unfortunately I again had some trouble convincing myself that all the intuitions were completely watertight, so I wrote out the following rather more pedantic account based on Le Cam's idea. For technical details see Problem [11].

Here is the key idea. Suppose $Y = (Y_1, Y_2, \ldots)$ is obtained by sampling repeatedly with replacement from U. With probability one each member of U appears infinitely often in the Y sequence. If we discard all except the first appearance of each u in U from the Y sequence then we are left with a random permutation, $(X_1, \ldots, X_N)$ of U; and $X_1, \ldots, X_n$ forms a sample of size n taken without replacement from U.

The sequence $(Y_1, Y_2, \ldots)$ will contain repeats, which can be represented as a sequence $\mathcal{C}(Y) = (\mathcal{C}_1(Y), \mathcal{C}_2(Y), \ldots)$ of symbols from a set of 'code-words' $\mathbb{B} = \{\boxed{j} : 1 \le j \le N\}$, by the following procedure. Think of $\mathbb{B}$ as ordered: $\boxed{1} < \boxed{2} < \cdots < \boxed{N}$. The code $\mathcal{C}(Y)$ always starts with $\boxed{1}$. If $Y_2 = Y_1$ then $\mathcal{C}_2(Y) = \boxed{1}$, otherwise $\mathcal{C}_2(Y) = \boxed{2}$. And so on. In general, if a Y_i repeats an earlier Y_j then $\mathcal{C}_i(Y) = \mathcal{C}_j(Y)$; if Y_i is different from all previous Y_j's then it receives the smallest unused code symbol. For example, here is how it works for a typical Y:

Y:	u_7	u_3	u_9	u_7	u_2	u_3	u_3	u_{185}	$\ldots$
X:	u_7	u_3	u_9		u_2			u_{185}	$\ldots$
$\mathcal{C}(Y)$:	$\boxed{1}$	$\boxed{2}$	$\boxed{3}$	$\boxed{1}$	$\boxed{4}$	$\boxed{2}$	$\boxed{2}$	$\boxed{5}$	$\ldots$

You should ignore the gaps in the X-vector; I inserted them just to align each X_j with its first appearance in the Y sequence. The corresponding positions in $\mathcal{C}(Y)$ contain a repetition of an earlier code symbol. For ex-

ample, the second u_7 in the Y sequence has a gap in X and is coded as $\boxed{1}$ because $Y_1 = u_7$.

Together, X and $\mathcal{C}(Y)$ allow us to reconstruct Y: for $i = 1, \ldots, N$ replace each $\boxed{i}$ in $\mathcal{C}(Y)$ by X_i, the ith element of X (ignoring the $\sqcup$ characters). More concisely, $Y_j = X_{\mathcal{C}(Y_j)}$, provided we ignore the little box around the code symbol.

I claim that X and $\mathcal{C}(Y)$ are independent.

Remark. Initially I thought the independence was obvious: knowledge of the pattern tells us nothing about the order in which the elements of U are first observed. For example, if $\mathcal{C}(Y) = (\boxed{1}, \boxed{2}, \boxed{3}, \boxed{1}, \ldots)$ then we know that Y_1, Y_2, Y_3 are different elements of U and $Y_4 = Y_1$ but we have no information about which three elements of U were involved. Then I began to worry that this assertion was a bit too handwaving. It took me a while to come up with the more rigorous argument given in Problem [11].

Now back to <48>. Remember that $S_n = g(Y_1) + \ldots g(Y_n)$ and $T_n = g(X_1) + \ldots g(X_n)$. The sum S_n can be re-expressed using the counts

$$N_n(j) = \text{number of times } \boxed{j} \text{ appears amongst } \mathcal{C}(Y_1), \ldots, \mathcal{C}(Y_n).$$

For example, if $n = 6$ and $Y = (u_7, u_3, u_9, u_7, u_2, u_3, \ldots)$ then

$$(X_1, \ldots, X_4) = (u_7, u_3, u_9, u_2) \quad \text{AND} \quad \mathcal{C}(Y) = (\boxed{1}, \boxed{2}, \boxed{3}, \boxed{1}, \boxed{4}, \boxed{2} \ldots),$$

so that $N_6(1) = N_6(2) = 2$ and $N_6(j) = 1$ for $j = 3, 4$, which gives

$$g(Y_1) + \cdots + g(Y_6) = 2g(X_1) + 2g(X_2) + g(X_3) + g(X_4).$$

Notice that we only need the counts up to $j = 6$, at most, because $(Y_1, \ldots, Y_6)$ can involve at most 6 different elements $X_1, \ldots, X_6$ of U. In general,

$$S_n = g(Y_1) + \cdots + g(Y_n) = \sum_{j=1}^{n} N_n(j) g(X_j)$$

and

\E@ Y.rep <49>
$$\mathbb{P}f\Big(S_n\Big) = \mathbb{P}f\left(\sum_{j=1}^{n} N_n(j) X_j\right)$$

Unfortunately, the final expression is not symmetric in $X_1, \ldots, X_n$; it is hard to see how it is related to $\mathbb{P}f(T_n)$. My method for determining patterns broke the symmetry but it can be restored using a sneaky trick. As the $N_n(j)$'s depend only on $\mathcal{C}(Y)$ they are independent of X. We could replace $(X_1, \ldots, X_n)$ by any other random sequence $(\widetilde{X}_1, \ldots, \widetilde{X}_n)$ that is independent of $\mathcal{C}(Y)$ and has the same distribution as $(X_1, \ldots, X_n)$. For example, for any fixed permutation σ of $[\![n]\!] := \{j \in \mathbb{N} : j \le n\}$ we could use $(X_{\sigma(1)}, \ldots, X_{\sigma(n)})$:

$$\mathbb{P}f\Big(S_n\Big) = \mathbb{P}f\left(\sum_{j=1}^{n} N_n(j) X_{\sigma(j)}\right) \qquad \text{for each permutations } \sigma \text{ of } [\![n]\!].$$

We can even average out over the uniform distribution $\mathbb{Q}$ on the set of all permutations of $[[n]]$ then use the JENSEN inequality to take the $\mathbb{Q}$ integral inside the convex function:

$$\mathbb{P}f\Big(S_n\Big) = \mathbb{Q}^\sigma \mathbb{P}f\left(\sum\nolimits_{j=1}^n N_n(j)X_{\sigma(j)}\right) \geq \mathbb{P}f\left(\mathbb{Q}^\sigma \sum\nolimits_{j=1}^n N_n(j)X_{\sigma(j)}\right)$$

From the facts that $\sum_{j=1}^n N_n(j) = n$ and

$$\mathbb{Q}^\sigma g(X_{\sigma(j)}) = n^{-1}\sum\nolimits_{i=1}^n g(X_i) = n^{-1}T_n \qquad \text{for } 1 \leq j \leq n$$

□ it now follows that $\mathbb{P}f(S_n) \geq \mathbb{P}f(T_n)$, as asserted.

3.9 Problems

MGF::S:Problems

For Problems [1] through [5], the function $\phi(x)$ denotes the $N(0,1)$ density and $\bar{\Phi}(x) := \int_x^\infty \phi(t)\,dt$; the functions $\mathcal{R}(\cdot)$ and $\rho(\cdot)$ and $r(\cdot)$ are defined on $\mathbb{R}$ by $1/\rho(x) := \mathcal{R}(x) = \bar{\Phi}(x)/\phi(x)$ and $r(x) := \rho(x) - x$.

MGF::P:Laplace [1] Inequality <10> is just the initial part of a sequence of upper and lower bounds for $\mathcal{R}(x)$, which are apparently due to Laplace (see Notes). Each bound is of the form $p(1/x)$ with p a polynomial.

(i) Show that $p(1/x) > \mathcal{R}(x)$ for all $x > 0$ if

\E@ Mill.upper <50> $$-\frac{d}{dt}\left(p(1/t)\phi(t)\right) > \phi(t) \qquad \text{for all } t > 0$$

and $p(1/x) < \mathcal{R}$ for all $x > 0$ if

\E@ Mill.lower <51> $$-\frac{d}{dt}\left(p(1/t)\phi(t)\right) < \phi(t) \qquad \text{for all } t > 0$$

Hint: $\int_x^\infty$.

(ii) Show that<50> holds if and only if $p(s) + s^3\dot{p}(s) > s$ for all $s > 0$. Characterize <51> by the reverse inequality.

(iii) Define a sequence of monomials by $\Delta_0(s) = s$ and $\Delta_k(s) = -s^3\dot{\Delta}_{k-1}(s)$ for $k \geq 1$. Define $a_k := 1 \times 3 \times \cdots \times (2k-1)$. Show that

$$\Delta_k(s) = (-1)^k a_k s^{2k+1} = -(2k-1)s^2\Delta_{k-1}(s) \qquad \text{for } k = 1, 2, \ldots.$$

(iv) Define $p_k(s) = \sum_{i=0}^k \Delta_i(s)$. Show that $p_k(s) + s^3\dot{p}_k(s) = s - \Delta_{k+1}(s)$.

(v) Conclude that $p_k(1/x) > \mathcal{R}(x) > p_{k+1}(1/x)$ for each even k.

MGF::P:half [2] Show that $\bar{\Phi}(x) \leq \frac{1}{2}e^{-x^2/2}$ for $x \geq 0$. Hint: From Theorem <19>(i) we have $\mathcal{R}(0) > \mathcal{R}(x)$.

MGF::P:Birnbaum [3] (Birnbaum, 1942). Use the CAUCHY-SCHWARZ inequality and some results from Theorem <19> to show that

$$\phi(x)^2 = \left(\int_x^\infty t\sqrt{\phi(t)}\sqrt{\phi(t)}\,dt\right)^2 \leq \big(x\phi(x) + \bar{\Phi}(x)\big)\,\bar{\Phi}(x).$$

Deduce that $\rho(x)^2 \le x\rho(x)+1$, so that $|\rho(x)-x/2| \le \sqrt{1+x^2/4}$, which implies $\rho(x) \le \left(x+\sqrt{x^2+4}\right)/2$.

MGF::P:Sampford

[4] (Sampford, 1953) Define $\psi(x) := 2r(x)^2 + xr(x)$, as in the proof of Theorem <19>, which showed that $\psi(x) > 1$ for all $x \in \mathbb{R}$. Argue that $r(x)$ cannot belong to the closed interval $I_x := \{t \in \mathbb{R} : 2t^2 + xt - 1 \le 0\}$, which has endpoints $\left(-x \pm \sqrt{x^2+8}\right)/4$. Deduce that $r(x) > (-x+\sqrt{x^2+8})/4 = 2/(x+\sqrt{x^2+8})$ and $\rho(x) > (3x+\sqrt{x^2+8})/4$. Note: $r(x) > 0$.

MGF::P:expected.max

[5] Suppose $Z_1, \ldots, Z_n$ are random variables, each distributed $N(0,1)$ but, for the moment, not necessarily independent. Define $M_n = \max_{i\le n} Z_i$.

(i) Even without independence the MGF approach also gives an upper bound a_n for the expected value of M_n, via Jensen's inequality: for each $\lambda > 0$,

$$\exp(\lambda \mathbb{P} M_n) \le \mathbb{P} e^{\lambda M_n} = \mathbb{P} \max_i e^{\theta Z_i} \le \sum\nolimits_i \mathbb{P} e^{\lambda Z_i} = n e^{\lambda^2/2}.$$

Deduce that $\mathbb{P}M_n \le \inf_{\lambda>0} \left(\log n + \lambda^2/2\right)/\lambda = a_n = \sqrt{2\log(n)}$. The case where $Z_i = Z_1$ for all i shows that the bound is not sharp in general.

(ii) If the Z_i's are independent, show that $\mathbb{P}M_n \ge a_n - c\log(a_n)/a_n$ for some constant c, if n is large enough. First show that

$$M_n \ge \max_{i\le n} Z_i^+ - \left(\sum\nolimits_{i\le n} |Z_i|\right) \{M_n \le 0\}$$

so that $\mathbb{P}M_n \ge \mathbb{P}\max_{i\le n} Z_i^+ - n\mathbb{P}|Z_1|/2^{n-1}$. Then argue that

$$O(n/2^{n-1}) + \mathbb{P}M_n \ge \mathbb{P}\max_{i\le n} Z_i^+ = \int_0^\infty \mathbb{P}\{M_n > t\}\, dt \ge x_n \mathbb{P}\{M_n > x_n\}.$$

Look at Example <17> for a way to choose x_n.

MGF::P:max.abs.normals

[6] Suppose $Z_1, \ldots, Z_n$ are independent random variables, each distributed $N(0,1)$.

(i) Show that $\mathbb{P}\{\max_{i\le n} |Z_i| \le x_n\} = \left(1 - 2\bar{\Phi}(x_n)\right)^n$.

(ii) Mimic the argument from Example <17> to deduce that $\max_{i\le n}|Z_i|$ concentrates near a_n.

MGF::P:std.exp

[7] Suppose X has a standard exponential distribution.

(i) Show that $\mathbb{P}\{X \ge x\} = e^{-x}$ for all $x \ge 0$ and $\mathbb{P}X = 1$.

(ii) Show that the method from Section 3.1 gives $\mathbb{P}\{X \ge x\} \le (ex)e^{-x}$ for $x \ge 1$.

(iii) What bound does the method give for $0 \le x < 1$?

MGF::P:union.indep

[8] Suppose $A_1, \ldots, A_n$ are independent events with $\sum_i \mathbb{P}A_i = \epsilon$, for a small ϵ. Show that

$$\mathbb{P} \cup_i A_i = 1 - \exp\left[\sum\nolimits_i \log(1 - \mathbb{P}A_i)\right] \ge 1 - e^{-\epsilon} = \epsilon - O(\epsilon^2).$$

MGF::P:Stein.exch [9] Use notation from Example <28>. Suppose W has distribution P. Suppose we can construct a new random variable $\widetilde{W}$ such that the joint distribution of $(W, \widetilde{W})$ is the same as the joint distribution of $(\widetilde{W}, W)$ (an "exchangeable pair"). Write Δ for $\widetilde{W} - W$ and define

$$\mu_1(w) := \mathbb{P}(\Delta \mid W = w) \quad \text{AND} \quad \mu_2(w) := \mathbb{P}\left(\Delta^2 \mid W = w\right).$$

Assume that $\|h\|_{Lip}$ is finite.

(i) For real x and $z := y - x$ define an antisymmetric function

$$F(x, y) := (y - x)\,[f(x) + f(y)] = z\left[2f(x) + z\dot{f}(x)\right] + R(x, z),$$

where $R(x, z) := f(x + z) - f(x) - z\dot{f}(x)$, as in <31>. Show that

$$0 = \mathbb{P}F(W, \widetilde{W}) = P^w\left(2\mu_1(w)f(w) + \mu_2(w)\dot{f}(w)\right) + \mathbb{P}R(W, \Delta).$$

(ii) Suppose $\mu_1(w) = -\lambda w + R_1(w)$ and $\mu_2(w) = 2\lambda + R_2(w)$ for some positive constant λ, with R_1 and R_2 small. Deduce from (i) that

$$2\lambda|\mathbb{P}H(W)| \le 2\|h\|_{Lip}\mathbb{P}|\Delta|^3 + \sqrt{\pi/2}\|H\|_\infty\mathbb{P}|R_1(W)| + 2\|H\|_\infty\mathbb{P}|R_2(W)|.$$

Compare with Stein (1986, pp. 33–35).

MGF::P:CLT.nsc [10] Let $\{\xi_{n,i} : i \in [\![n]\!],\ n \in \mathbb{N}\}$ be a triangular array of random variables, independent within each row. It is a classical result (see Petrov, 1975, §IV.4 and Le Cam, 1986b) that if $\max_i \mathbb{P}\{|\xi_{n,i}| > \epsilon\} \to 0$ for each $\epsilon > 0$ then $\sum_i \xi_{n,i} \rightsquigarrow N(0,1)$ iff for each $\epsilon > 0$ we have: $\mathbb{P}\{\max_i |\xi_{n,i}| > \epsilon\} \to 0$ and $\sum_i \operatorname{var}(\xi_{n,i}\{|\xi_{n,i}| \le \epsilon\}) \to 1$ and $\sum_i \operatorname{var}(\xi_{n,i}\{|\xi_{n,i}| \le \epsilon\}) \to 1$. Thus, for the purpose of developing normal approximations for sums of independent random variables it suffices to consider the case where $\xi_1, \ldots, \xi_n$ are independent random variables with: $|\xi_i| \le \epsilon$ for a fixed $\epsilon > 0$; $\mathbb{P}\xi_i = 0$; and $\sum_i \operatorname{var}(\xi_i) = 1$. This Problem will show how the bounds from Problem [9] can handle such a situation.

Let $\xi_1, \ldots, n, \eta_1, \ldots, \eta_n, J$ be independent random variables, with η_i having the same distribution as ξ_i and J being uniformly distributed on $[\![n]\!]$. Define $W := \sum_i \xi_i$ and $\widetilde{W} = \sum_i \{J \ne i\}\xi_i + \{J = i\}\eta_i$, so that $\Delta := \widetilde{W} - W = \sum_i \{J = i\}(\eta_i - \xi_i)$.

(i) Show that $\mu_1(w) = -w/n$ and $\mu_2(w) = 2n^{-1} + R_2(w)$ where $R_2(w) = n^{-1}P_w\sum_i\left(\xi_i^2 - \sigma_i^2\right)$, where $\sigma_i^2 := \mathbb{P}\xi_i^2$ and P_w is shorthand for the conditional distribution of $\widetilde{W}$ given $W = w$.

(ii) Show that

$$(n\mathbb{P}|R_2(W)|)^2 \le \mathbb{P}\left(\sum\nolimits_i \xi_i^2 - \sigma_i^2\right)^2 \le \sum\nolimits_i \mathbb{P}\xi_i^4 \le \epsilon^2,$$
$$n\mathbb{P}|\Delta|^3 \le |\eta_i - \xi_i|^3 \le 4\epsilon.$$

(iii) Deduce that $|\mathbb{P}H(W)| \le C\max(\|h\|_{Lip}, \|H\|_\infty)\epsilon$ for some universal constant C.

Remark. The previous Problem is just a sanity check for Problem [9]. The real strength of Stein's approach is its ability to handle problems caused by dependence, which defy classical method.

MGF::P:indep.code

[11] Here is a rigorous way to establish independence of X and $\mathcal{C}(Y)$ in Example <47>. Notation (derived from Section 3.8):

a) Regard Y as the identity map (that is, $Y(y) = y$) on $U^{\mathbb{N}}$ equipped with its product sigma-field and product measure $\mathbb{P} = \nu^{\mathbb{N}}$, where ν denotes the uniform distribution on U.

b) $\mathbb{B} = \{\boxed{i} : i = 1, \ldots, N\}$, the code symbols. Regard $\mathcal{C}$ as a measurable map from $U^{\mathbb{N}}$ into the product space $\mathbb{B}^{\mathbb{N}}$ (equipped with its product sigma-field).

c) $\mathcal{W}$ = the set of all permutations of U. (Thus $|\mathcal{W}| = N!$.) If $x = (x_1, \ldots, x_N) \in \mathcal{W}$ and $\boxed{i} \in \mathbb{B}$, interpret $x_{\boxed{i}}$ to mean x_i.

d) Treat $X = X(y)$ as the map from $U^{\mathbb{N}}$ into the set $\mathcal{W}$ that is defined by discarding repetitions of each y_i after its first appearance in y.

e) For $y = (y_1, y_2, \ldots) \in U^{\mathbb{N}}$ and $B = (b_1, b_2, \ldots) \in \mathbb{B}^{\mathbb{N}}$ define $\mathcal{C}(y) \,[\!]\, n = (\mathcal{C}_1(y), \ldots, \mathcal{C}_n(y))$ and $B \,[\!]\, n = (b_1, \ldots, b_n)$.

The argument:

(i) Show that the distribution of X is uniform on $\mathcal{W}$. Deduce that

$$\mathbb{P}\{X_j = x_j, \ldots, X_N = x_N\} = \frac{1}{N} \times \frac{1}{N-1} \times \frac{1}{N-j+1} = \frac{1}{(N)_j}$$

for each $x \in \mathcal{W}$ and $1 \le j \le N$. (Exchangeability helps.)

(ii) Show that the distribution of $\mathcal{C}(Y)$ concentrates on the set $\mathcal{P}$ of all *feasible* B's in $\mathbb{B}^{\mathbb{N}}$, that is, those B that start with $\boxed{1}$ and for all i, j with $1 \le j < i \le N$ the codeword $\boxed{j}$ first appears in B before $\boxed{i}$.

(iii) Suppose $B \in \mathcal{P}$ and $B \,[\!]\, n$ uses only the k code symbols $\boxed{1}, \ldots, \boxed{k}$. Show that

$$\mathbb{P}\{\mathcal{C}(Y) \,[\!]\, n = B \,[\!]\, n\} = \frac{N(N-1)\ldots(N-k+1)}{N^n} = \frac{(N)_k}{N^n}.$$

Hint: Each repeat of a codeword corresponds to an event that has probability $1/N$ and the first appearance of codeword $\boxed{j}$ indicates a selection from a set $N - j + 1$ elements from U.

(iv) Suppose $x = (x_1, \ldots, x_N) \in \mathcal{W}$ and $B \in \mathcal{P}$, with $B \,[\!]\, n$ using only the k code symbols $\boxed{1}, \ldots, \boxed{k}$. Justify the following assertions.

Define

$$A = \{y : \mathcal{C}(y) \,[\!]\, n = B \,[\!]\, n, \quad X_i = x_i \text{ for } i = 1, \ldots, k\}.$$

The assumption about B means that $y_1, \ldots, y_n$ select only k distinct elements of U, namely $U_1 := \{x_i : i = 1, \ldots, k\}$, with first selections occurring in the order $(x_1, \ldots, x_k)$. Thus

$$\mathbb{P}A = \mathbb{P}\{y : y_i = x_{b_i} \text{ for } i = 1, \ldots, n \} = (1/N)^n.$$

Conditional on the occurrence of A, the remaining observations $y_{n+1}, y_{n+2}, \ldots$ are left to select each element of $U_2 := \{x_i : n + 1 \le i \le N\}$. If we also require $X_i = x_i$ for $k + 1 \le i \le N$ then we have specified the order of first selections of the elements of U_2, namely $(x_{k+1}, \ldots, x_N)$. Thus

$$\mathbb{P}\{X_i = x_i \text{ for } k + 1 \le i \le N \mid A\} = \frac{1}{N-k} \times \frac{1}{N-k-1} \times \cdots \times \frac{1}{1} = \frac{1}{(N-k)!}.$$

Combining these two results we get

$$\begin{aligned}
&\mathbb{P}\{y : \mathcal{C}(y) \,[\!]\, n = B \,[\!]\, n,\ X = (x_1, \ldots, x_N)\} \\
&\quad = \mathbb{P}\left(A \cap \{X_i = x_i \text{ for } k + 1 \le i \le N \}\right) \\
&\quad = (1/N)^n \times \frac{1}{(N-k)!} = \frac{N(N-1)\ldots(N-k+1)}{N^n} \times \frac{1}{N!} \\
&= \mathbb{P}\{\mathcal{C}(Y) \,[\!]\, n = B \,[\!]\, n\} \times \mathbb{P}\{X = x\}.
\end{aligned}$$

It follows that $\mathcal{C}(Y)$ and X are independent.

3.10 Notes

MGF::S:Notes

Bennett (1962) and Hoeffding (1963) are good sources for a host of exponential inequalities. Massart (2003, Chapter 2) and Boucheron, Lugosi, and Massart (2013, Chapter 2) persuaded me that it is a good idea to have the relevant ideas collected together in one place, rather than deriving them on an ad hoc basis.

Many authors seem to credit Chernoff (1952) with the moment generating trick in <1>, even though the idea is obviously much older. In that paper Chernoff first noted that Cramér (1938) had already established excellent results for sums of independent random variables using the MGF method. Then he proceeded (page 495) to list a few basic facts:

> Since the results of Cramér are extremely more powerful tha[n] we require here and the (finite) existence of third order moments is not necessary for the results that we desire, we shall state and briefly outline a proof of Theorem 1. Before doing this we shall first formally state some notation and lemmas which we shall use throughout this paper. These lemmas state known results which are rather obvious, depending mainly on Lebesgue's Theorem on integration of monotone sequences [reference to the Saks book].

Remark. Cramér's 1938 paper summarized asymptotic approximations to the tail probabilities for a sum of independent random variables, rather than bounds on those tail probabilities; details appeared in Chapter 7 of Cramér (1937). See Cramér (1976, Section 4.9) for comments about the conference where he presented the 1938 paper.

Amongst the known results listed by Chernoff was a special case of the 'extended Tchebycheff' inequality

$$\mathbb{P}\{S \geq x\} \leq \inf_{\lambda \geq 0} e^{-\lambda x} \mathbb{P} e^{\lambda S},$$

for which he cited the German original of Kolmogorov (1933). (The inequality appears in §IV.3.) In a subsequent biographical article Chernoff (2004) acknowledged that while working on his 1952 paper he had originally been ignorant of the 1938 Cramér paper.

In the 1933 book Kolmogorov gave no source for the 'extended Tchebycheff' inequality, although the 1927 edition of Sergei Bernstein's probability book was listed in his Bibliography. Moreover, in a paper celebrating Bernstein's eightieth birthday, Kolmogorov and Sarmanov (1960) noted:

> 3. Beginning in 1921 Sergei Natanovich published a number of papers dealing with various special problems in the application of probability theory ... and in 1927 appeared the first edition of the fundamental text "The Theory of Probability", which was reprinted with large supplements in 1934 and 1946. At the mathematical congresses in Moscow (1927) and Zürich (1932) Sergei Natanovich delivered long survey reports on the problems of probability theory. We ... emphasize that at this time such a wide range of work on all the fundamental theoretical and applied problems of probability theory was a totally new thing. ... It is natural that the theoretical and applied works of Sergei Natanovich and his text in probability theory have determined to a considerable degree the development of research in probability theory in the USSR.

And then

> 4. A whole series of papers by Sergei Natanovich are connected with the strengthening of Chebyshev's inequality [citing papers from 1918, 1924, and 1937] and the calculation of the error in the Laplace formula ...

I took the quotes from the SIAM translation of Volume V number 2 of the Russian original.

Now for the view from the West. The proof of the Bernstein inequality given by Uspensky (1937, pages 204–206) used the MGF method. He prefaced the "Indication of the Proof" by the remark "S. Bernstein has shown that Tchebycheff's inequality can be considerably improved". I thank Elena Khusainova, who translated four pages from Bernstein's probability book for

me, clearly showing that Uspensky was following that book, which appeared in his list of references (page 207).

Hoeffding (1963, page 14) gave Bernstein credit for the MGF approach:

> The method employed to derive the inequalities, which has often been used (apparently first by S. N, Bernstein),...

Bennett (1962) had also cited (page 34) Bernstein's 1927 book (and an earier paper) with the comment (page 35):

> Bernstein's original work was published in Russian, and appears to be unobtainable. It is reported—indirectly—by [Craig, 1933] ...and by [Uspensky (1937, pages 204–206)] who indicates the proof in a series of exercises. The inequality is mentioned or quoted without proof by ...Apart from these brief references, Bernstein's inequality seems to have escaped notice in the English-speaking world.

Craig's had commented (his page 94):

> Another interesting and important attempt in this direction due to S. N. Bernstein seems to have generally escaped attention in the English-speaking world, at least, since it has been published only in Russian.

with the footnote

> "Bernstein, S., Theory of Probability, (Moscow, 1927), pp. 159-165. The present account of this work of Bernstein is taken from a lecture of Professor J. V. Uspensky."

Craig also mentioned that his paper was written while he was at Stanford University, where Uspensky was a mathematics faculty member, until his death in 1947.

I think it abundantly clear that credit for the MGF method should go to Bernstein, not Chernoff.

The result derived in Problem [1] corresponds to an analogous result for the error function (that is, $\int_0^x e^{-t^2} dt$) presented by Laplace in his Celestial Mechanics, reprinted 1805 in Volume IV, Book X, Chapter 1, §5 of his collected works (pages 489–492 in the Bowditch translation). He also developed a continued fraction expansion.

The ratio $\mathcal{R}(x)$ is often called the "Mills ratio", because it was discussed by Mills (1926). Actually Mills was just constructing a table of $\mathcal{R}$, using earlier tables for the normal distribution function and numerical methods proposed by other authors. There has been a long history of authors deriving upper and lower bounds for $\mathcal{R}$, such as the upper bound from Problem [3] and the lower bound from Problem [4]. See Baricz (2008) or Gasull and Utzet (2014) for recent examples, which include some history. I first came to appreciate the value of having global control over ρ while working out the

theory for Carter and Pollard (2004) and UGMTP Appendix D. Most of the proof for Theorem <19> comes from those two sources, with corrections, plus ideas adapted from Sampford (1953).

The subGamma idea is clearly present in the derivation of Bernstein's inequality from the bernstein moment assumption (see Section 8.3) but, to my knowledge, Boucheron, Lugosi, and Massart (2013, page 28) were the first to anoint it as a general concept. It seems that the neat trick <39> with the square roots was first noted by Birgé and Massart (1998, Section 7.6).

Okamoto (1959) stated the binomial tail bounds <43> and its analog for the lower tail in a slightly different form, with the comment that "We shall state two Lemmas the first of which is a corollary of a theorem given by [Chernoff 1952, Theorem 1]". He then derived several more attractive bounds that could be derived from the MGF bound. He omitted the calculus (which I provided in Section 3.7) for the MGF bound; he only gave the details for the weaker upper bounds. He also commented that his proof simplified a "tedious, although elementary" calculation by Uspensky (1937, page 102). It seems strange to me that Okamoto did not also cite Uspensky (1937, page 204).

References

Baricz2008mills

Baricz, Á. (2008). Mills' ratio: monotonicity patterns and functional inequalities. *Journal of Mathematical Analysis and Applications 340*(2), 1362–1370.

Bennett62jasa

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

BirgeMassart98Bernouilli

Birgé, L. and P. Massart (1998). Minimum contrast estimation on sieves: exponential bounds and rates of convergence. *Bernouilli 4*(3), 329–375.

Birnbaum1942AMS

Birnbaum, Z. W. (1942). An inequality for Mills' ratio. *Ann. Math. Statist. 13*, 245–246.

BLM2013Concentration

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

CarterPollard04tusnady

Carter, A. and D. Pollard (2004). Tusnády's inequality revisited. *Annals of Statistics 32*(6), 2731–2741.

Chernoff52AMS

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations. *Annals of Mathematical Statistics 23*(4), 493–507.

Chernoff2004Rubin

Chernoff, H. (2004). Some reminiscences of my friendship with Herman Rubin. In A. DasGupta (Ed.), *A Festschrift for Herman Rubin*, Volume 45

of *Lecture Note-Monograph Series*, pp. 1–4. Institute of Mathematical Statistics.

Cochran52ams Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics 23*, 315–3455.

Craig1933AMS Craig, C. C. (1933). On the Tchebychef inequality of Bernstein. *The Annals of Mathematical Statistics 4*(2), 94–102.

Cramer37book Cramér, H. (1937). *Random Variables and Probability Distributions*. Cambridge University Press.

Cramer1938 Cramér, H. (1938). Sur un nouveau théorème-limite de la théorie des probabilités. In *Colloque consacré à la théorie des probabilités, Actualités scientifiques et industrielles 736*, pp. 2–23. Hermann & Cie, Paris. Available at arXiv:1802.05988v3 in the form of a TeXed version of the original paper in French paired with an English translation by Hugo Touchette. Reprinted in: H. Cramér, Collected Works, A. Martin-Löf (Ed.), Vol. II, Springer, Berlin, 1994, p. 895-913.

Cramer76 Cramér, H. (1976). Half a century with probability theory: some personal recollections. *Annals of Probability 4*, 509–546.

GasullUtzet2014JMAA Gasull, A. and F. Utzet (2014). Approximating Mills ratio. *Journal of Mathematical Analysis and Applications 420*(2), 1832–1853.

HallBarbour1984PAMS Hall, P. and A. D. Barbour (1984). Reversing the Berry-Esseen inequality. *Proceedings of the American Mathematical Society 90*(1), 107–110.

HoChen1978AnnProb Ho, S.-T. and L. H. Y. Chen (1978). An L_p bound for the remainder in a combinatorial central limit theorem. *Annals of Probability 6*(2), 231 – 249.

Hoeffding1963JASA Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*, 13–30.

Kolmogorov33book Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer-Verlag. Second English Edition, *Foundations of Probability* 1950, published by Chelsea, New York.

KolmogorovSarmanov60 Kolmogorov, A. N. and O. V. Sarmanov (1960). The work of S. N. Bernshtein on the theory of probability (on his eightieth birthday). *Theory of Probability and Its Applications 5*, 197–203.

LaurentMassart2000AnnStat Laurent, B. and P. Massart (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics 28*(5), 1302–1338.

LeCam:86book Le Cam, L. (1986a). *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag.

LeCam1986statsci Le Cam, L. (1986b). The central limit theorem around 1935. *Statistical Science 1*, 78–96.

eadbetterLindgrenRootzen83 Leadbetter, M. R., G. Lindgren, and H. Rootzén (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag.

Massart03Flour Massart, P. (2003). *Concentration Inequalities and Model Selection*, Volume 1896 of *Lecture Notes in Mathematics*. Springer Verlag. Lectures given at the 33rd Probability Summer School in Saint-Flour.

Mills1926Biometrika Mills, J. P. (1926). Table of the ratio: area to bounding ordinate, for any portion of normal curve. *Biometrika 18*, 395–400.

Okamoto1959AnnInst Okamoto, M. (1959). Some inequalities relating to the partial sum of binomial probabilities. *Annals of the institute of Statistical Mathematics 10*(1), 29–35.

Pearson1900chi Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50*(302), 157–175.

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

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

Sampford1953AMS Sampford, M. R. (1953). Some inequalities on Mill's ratio and related functions. *The Annals of Mathematical Statistics 24*(1), 130–132.

Stein1972Berk6 Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. Sixth Berkeley Symp. Math. Stat. Prob.*, Volume 2, pp. 583–602.

Stein86ims Stein, C. (1986). *Approximate Computation of Expectations*, Volume 7 of *Lecture Notes–Monograph series*. Institute of Mathematical Statistics.

Uspensky1937book Uspensky, J. V. (1937). *Introduction to Mathematical Probability*. McGraw-Hill.

WuZhou2019arxiv 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, arXiv 1908.10935.

