9.07 Introduction to Statistics for Brain and Cognitive Sciences
Emery N. Brown

Lecture 7 Limit Theorems: Law of Large Numbers and the Central Limit Theorem

I. Objectives
Understand the logic behind the Law of Large Numbers and its relation to the frequency interpretation of probability theory

Understand how to prove the Weak Law of Large Numbers

Understand the logic behind the Central Limit Theorem

Understand how to prove a version of the Central Limit Theorem

Understand how to construct approximations to the distribution of the sums of random variables using the Central Limit Theorem

Understand the conditions under which the Gaussian distribution can be used to approximate binomial and Poisson probabilities

II. Law of Large Numbers
To motivate our study of the Law of Large Numbers we begin with an example.

Example 7.1. An Opinion Poll. An election is to be conducted for President of the Society for Neuroscience and there are two candidates for the office: Candidate A and Candidate B. If we poll n voters and we assume that each individual polled reports truthfully his/her voting intention then how sure can we be that the fraction of people who report that they will vote for Candidate A is a good "guess" (estimate) of the number who will actually vote for Candidate A?

Our intuition is that the fraction of people who say they are going to vote for Candidate A should be a good "guess" of the proportion of votes that Candidate A will get in the election. Furthermore, the more people we poll, the better our "guess" should be. With what we know already, we can formalize our intuition. If we let p denote the fraction of people who will vote for Candidate A, we can define X_i to be the Bernoulli random variable which is 1 if person i is going to vote for Candidate A and 0 otherwise. Let $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$. $\bar{X}$ is the fraction of people polled who vote for Candidate A. Our intuition is that $\bar{X}$ is a good guess for p. This intuition is what we colloquially call the "Law of Averages."

To begin thinking formally about this problem we recall that a **statistic** is any function of a set of data. Because experimental data are random and the statistic is a function of the data, the statistic is also a random variable. It therefore, has a probability density or probability mass function. If we observe random variables $X_1, X_2, ..., X_n$ such that the X_i's are independent and all have the same probability distribution, then the collection $X_1, X_2, ..., X_n$ is called a **random sample** or simply a **sample**. The sample mean is one of the most basic statistics and we recall that it is defined as

$$\bar{X} = n^{-1}\sum_{i=1}^{n} X_i. \tag{7.1}$$

Remark 7.1. If the mean and variance of the X_i's are respectively μ and variance σ^2, then by **Propositions 6.1** and **6.2** we have

$$E(\bar{X}) = n^{-1}\sum_{i=1}^{n} E(X_i) = \mu \tag{7.2}$$

$$Var(\bar{X}) = \sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n} \tag{7.3}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{n^{\frac{1}{2}}}. \tag{7.4}$$

The Law of Large Numbers tells us that as the sample size increases the probability that $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ is close to the mean μ approaches 1. To prove this, we establish first two technical results.

A. Some Technical Results

Proposition 7.1 (Markov Inequality). If X is a random variable that takes on only non-negative values, then for any $a > 0$

$$\Pr(X \geq a) \leq \frac{E[X]}{a}. \tag{7.5}$$

Proof: To see this note that

$$\begin{aligned} E[X] &= \int_0^\infty xf(x)dx = \int_0^a xf(x)dx + \int_a^\infty xf(x)dx \\ &\geq \int_a^\infty xf(x)dx \geq a\int_a^\infty f(x)dx = a\Pr(x \geq a). \end{aligned} \tag{7.6}$$

Proposition 7.2 (Chebyshev's Inequality). If X is a random variable with mean μ and variance σ^2 and $k > 0$, then

$$\Pr\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}. \tag{7.7}$$

Proof: Let $Y = |X - \mu|$ then $Y > 0$ and by the Markov Inequality

$$\Pr\{|X - \mu| \geq k\} = \Pr\{Y \geq k\} = \Pr\{Y^2 \geq k^2\} \leq \frac{E(Y^2)}{k^2} = \frac{\sigma^2}{k^2} \tag{7.8}$$

Having established Chebyshev's Inequality we can prove the Weak Law of Large Numbers.

B. The Law of Large Numbers

Proposition 7.3 Law of Large Numbers (Weak). Let $X_1, X_2, ..., X_n$ be an independent, identically distributed (i.i.d.) sample from a population with mean μ and variance σ^2. Then the probability that the difference between the sample mean and true mean remains greater than any finite amount goes to zero as the sample size n goes to infinity.

Proof: If we pick $\varepsilon > 0$, then we can

$$\Pr(|\bar{X} - \mu| < \varepsilon) = 1 - \Pr(|\bar{X} - \mu| > \varepsilon). \tag{7.9}$$

Using Chebyshev's Inequality, for $\varepsilon > 0$,

$$\Pr(|\bar{X} - \mu| > \varepsilon) \leq \frac{Var(\bar{X})}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2 n} \longrightarrow 0 \tag{7.10}$$

as n goes to infinity. Hence, $\Pr(|\bar{X} - \mu| < \varepsilon) = 1 - \Pr(|\bar{X} - \mu| > \varepsilon) \longrightarrow 1.$

Example 7.2 The Law of Large Numbers and the Behavior of the Sample Mean. To gain intuition about what is occurring with the Law of Large Numbers, we consider (**Figure 7A**) the mean of n observations from a gamma distribution with $\alpha = 2$ and $\beta = 1$ (**Figure 7A ,column 1**) and an arbitrary distribution (**Figure 7A, column 2**) for $n = 1, 2, 4, 16$ and 400. We can see that in the case of the gamma distribution (Figure 7A, column 1) as n grows, the distribution of the sample mean gets more and more tightly clustered around the true mean. Similarly, for the arbitrary distribution an analogous phenomenon is observed (**Figure 7A, column 2**).

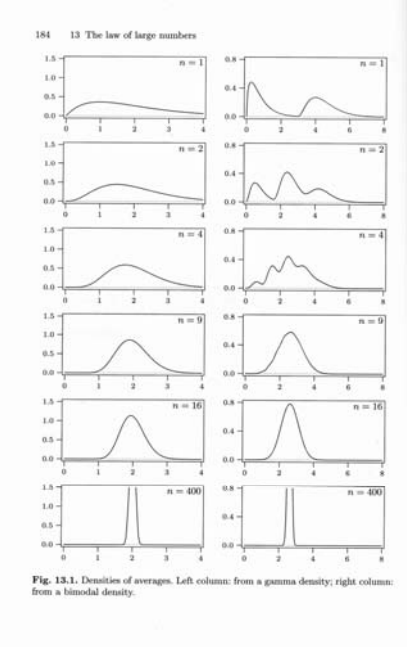

Fig. 13.1. Densities of averages. Left column: from a gamma density; right column: from a bimodal density.

Figure 7A. Illustration of the Law of Large Number by averaging n observations from a gamma distribution with $\alpha = 2$ and $\beta = 1$ (column 1) and a bimodal distribution (column 2) for $n = 1, 2, 4, 16$ and 400. Reproduced from Dekking et al. (2002).

Example 7.1 (continued). Can we use Chebyshev's Inequality to figure out how large n should be such that we are 95% certain that the difference between $\bar{X}$ and μ is less than 0.005? We will answer this question in **Homework Assignment 6**.

Remark 7.2. From the result in **Proposition 7.3**, we say that $\bar{X}$ converges in probability to μ. The best result known is the **Strong Law of Large Numbers**. It states that

$$\Pr(|\bar{X} - \mu| \to 0) = 1. \tag{7.11}$$

as long as μ exists. This means $\bar{X}$ is guaranteed to converge to μ in the usual numerical sense. Therefore, we are correct in thinking that an observed $\bar{X}$ is close to the true mean. This is not guaranteed by convergence in probability. The statement, "the sample mean is close to the true mean" does not apply to a particular realization. In contrast, the **Strong Law of Large Numbers** addresses what occurs to the result of a single realization of the data (Pawitan, 2001, pp. 231-232).

Remark 7.3. What is apparent from Figure 7A is that the sample mean is getting closer to the true mean. This is a first-order effect because it involves the mean (first moment). As the number of samples became large in Figure 7A, the shape of the distribution became more symmetric or "Gaussian-like". We might think of this effect as a second-order effect because it tells us something about the variability (second moment) of the statistic or the shape of its distribution. The second-order effect addresses the question of what is the behavior of the random variable $\bar{X}$ as it approaches its mean? That is, can we characterize its probability density? The Central Limit Theorem allows us to answer these questions.

III. The Central Limit Theorem

A. Motivation

The Central Limit Theorem is the most important theorem in probability theory and concerns the behavior of the sample mean as n grows indefinitely large. Before stating the Central Limit Theorem, we give an example to motivate its derivation along with some technical results we will need to prove it.

Example 7.3 A Fast Spiking Neuron. Consider a neuron with a constant rate λ that obeys a Poisson probability law. Assume that λ = 40 Hz. In a 1-second interval how likely are we to observe 40, 41, or 42 spikes?

We first develop a Gaussian approximation to compute this result then we compare it with the exact result computed directly from the Poisson probability mass function with λ = 40 Hz. To construct a Gaussian approximation we study the behavior of a sequence of Poisson random variables such that $X_n \sim Poisson(\lambda_n)$ .and $\lambda_n \to \infty$. We have that $E(X_n) = Var(X_n) = \lambda_n$, and hence, $E(X_n)$ and $Var(X_n) \to \infty$. Therefore, we must standardize X_n to get a limiting distribution. Let

$$Z_n = \frac{X_n - E(X_n)}{[Var(X_n)]^{\frac{1}{2}}} \tag{7.12}$$

$$Z_n = \frac{X_n - \lambda_n}{\lambda_n^{\frac{1}{2}}} \tag{7.13}$$

We have $E(Z_n) = 0$ and $Var(Z_n) = 1$. We show that $M_{Z_n}(t) \to e^{t^2/2}$ which by the Continuity Theorem (**Proposition 7.4,** stated below) means that $F_{Z_n}(z_n) \to F(z) = \Phi(z)$. Recall from **Lecture 6** (**Result 6.7** and **Remark 6.10**) that if X has mgf $M_X(t)$ and $Y = a + bX$, then $M_y(t) = e^{at} M_X(bt)$. We have $M_{X_n}(t) = e^{\lambda_n(e^t - 1)}$ and hence,

$$Z_n = \lambda_n^{-\frac{1}{2}} X_n - \lambda_n^{\frac{1}{2}},$$

$a = -\lambda_n^{\frac{1}{2}}$ and $b = \lambda_n^{-\frac{1}{2}}$

$$\begin{aligned} M_{Z_n}(t) &= e^{-\lambda_n^{\frac{1}{2}} t} M_{X_n}(\lambda_n^{-\frac{1}{2}} t) \\ &= e^{-t\lambda_n^{\frac{1}{2}}} e^{\lambda_n(e^{t\lambda_n^{-\frac{1}{2}}} - 1)}. \end{aligned} \tag{7.14}$$

We have that

$$\log M_{Z_n}(t) = -t\lambda_n^{\frac{1}{2}} + \lambda_n(e^{t\lambda_n^{-\frac{1}{2}}} - 1). \tag{7.15}$$

Recall that

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \ldots \tag{7.16}$$

so that Eq. 7.15 can be written as

$$\begin{aligned} \log M_{Z_n}(t) &= -t\lambda_n^{\frac{1}{2}} + \lambda_n\left(1 + \frac{t}{\lambda_n^{\frac{1}{2}}} + \frac{t^2}{2\lambda_n} + \frac{t^3}{3!\lambda_n^{\frac{3}{2}}} + \ldots + \ldots - 1\right) \\ &= -t\lambda_n^{\frac{1}{2}} + t\lambda_n^{\frac{1}{2}} + \frac{t^2}{2} + \frac{t^3}{3!\lambda_n^{\frac{1}{2}}} + \ldots + \ldots \\ &= \frac{t^2}{2} + \frac{t^3}{6\lambda_n^{\frac{1}{2}}} + \ldots + \ldots \end{aligned} \tag{7.17}$$

As $\lambda_n \to \infty, \log M_{z_n}(t) \to \frac{t^2}{2}$ and thus $M_{z_n}(t) \to e^{\frac{t^2}{2}}$. By the **Continuity Theorem** stated below, we conclude that $F_{z_n}(z_n) \to F_z(z) = \Phi(z)$.

In Homework Assignment 6, we present an alternative derivation of this result by showing directly that as $\lambda_n \to \infty$ the Poisson pmf $P_{z_n}(z_n) \to \phi(z)$ the pdf of a standard Gaussian random variable.

Applying this result to our fast spiking neuron in **Example 7.3** we have

Gaussian Approximation

$$\begin{aligned}\Pr(40 \le X \le 42) &= \Pr\left(\frac{39.5-\lambda}{\sqrt{\lambda}} \le \frac{X-\lambda}{\sqrt{\lambda}} \le \frac{42.5-\lambda}{\sqrt{\lambda}}\right)\\ &= \Pr\left(\frac{39.5-40}{\sqrt{40}} \le z \le \frac{42.5-40}{\sqrt{40}}\right)\\ &= \Phi(0.395) - \Phi(-0.079)\\ &= 0.6537 - 0.4685 = 0.1852,\end{aligned} \tag{7.18}$$

where we have used the continuity correction defined below in section **III E**.

Exact Poisson Solution

$$\Pr(X = 40, 41 \text{ or } 42) = \sum_{k=40}^{42} \frac{40^k e^{-40}}{k!} = 0.1828. \tag{7.19}$$

We see that in this case, the approximation is very reasonable.

B. Proof of the Central Limit Theorem

To prove the **Central Limit Theorem**, we need to state first a definition, a theorem and a technical result.

Definition 7.1. If $X_1, ..., X_n$ is a sequence of random variables with cumulative distribution functions $F_1, F_2, ...$ and let X be a random variable with distribution function F. The sequence X_n converges in distribution to X if

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} \Pr(X_n \le x) = \Pr(X \le x) = F(x). \tag{7.20}$$

Proposition 7.4 (Continuity Theorem). Let F_n be a sequence of cumulative distribution functions with corresponding moment generating functions $M_n(t)$. Let $F(X)$ be a cumulative distribution function with moment generating function $M(t)$. If $M_n(t) \to M(t)$ for all t in an open interval containing zero, then $F_n(X) = F(X)$ at all continuity points of $F(X)$.

Result 7.1. If $a_n \to a,$ then

$$\lim_{n\to\infty}(1+\frac{a_n}{n})^n = e^a. \tag{7.21}$$

This limit is established in advanced calculus courses.

Proposition 7.5. Central Limit Theorem: Let $X_1, X_2, ..., X_n$ be a sequence of independent random variables having mean μ and variance σ^2, common distribution function $F(x)$ and moment generating function $M(t)$ defined in a neighborhood of zero. Let $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ as $n \to \infty$

$$\Pr(\frac{n^{\frac{1}{2}}(\bar{X}-\mu)}{\sigma} \le x) \to \Phi(x). \tag{7.22}$$

Proof: To simplify the calculations, we assume $\mu = 0.$ Let

$$S_n = \sum_{i=1}^{n} X_i$$

and

$$Z_n = \frac{S_n}{\sigma n^{\frac{1}{2}}}. \tag{7.23}$$

Because the $X_1, ..., X_n$ are independent, the mgf of S_n is

$$M_{S_n}(t) = [M(t)]^n$$

and by **Result 6.7** and **Remark 6.10**

$$M_{Z_n}(t) = [M(\frac{t}{\sigma n^{\frac{1}{2}}})]^n. \tag{7.24}$$

The Taylor series expansion of $M(s)$ about 0 is

$$M(s) = M(0) + sM'(0) + \frac{1}{2}s^2M''(0) + \frac{s^3M'''(0)}{3} + ... + ... \tag{7.25}$$

Now $M(0) = 1,\ E(X) = 0,\ M'(0) = 0$ and $M''(0) = \sigma^2.$ We have

$$M(\frac{t}{\sigma n^{\frac{1}{2}}}) = [1 + \frac{1}{2}(\frac{t}{\sigma n^{\frac{1}{2}}})^2\sigma^2 + \frac{t^3}{3(\sigma n^{\frac{1}{2}})^3} + ... + ...$$
$$= 1 + \frac{t^2}{2n} + \frac{t^3}{3\sigma^3 n^{\frac{3}{2}}} + ... + ... \quad (7.26)$$

Note that

$$\frac{\frac{t^3}{3\sigma^3 n^{\frac{3}{2}}}}{\frac{t^2}{2n}} = \frac{2t}{3\sigma^3 n^{\frac{1}{2}}} \to 0 \quad (7.27)$$

as $n \to \infty$. The same is true for any term larger than t^3. Hence,

$$M_{z_n}(t) \approx [1 + \frac{t^2}{2n}]^n \quad (7.28)$$

for n large. By **Result 7.1** and **Proposition 7.4**

$$M_{z_n}(t) \to e^{\frac{t^2}{2}} \quad (7.29)$$

which is the mgf of a standard Gaussian random variable.

C. Applications of the Central Limit Theorem

Example 7.3 (continued). The Gaussian Approximation to the Poisson Distribution. In the case of the Poisson distribution with parameter λ we take the Gaussian mean and standard deviation to be $\mu = \lambda$ and $\sigma = \lambda^{\frac{1}{2}}$ respectively. This approximation is considered to be quite accurate according to Port (1994, p. 685) for $\lambda \geq 25$. That is, we must have on average 25 events per unit time for the Gaussian distribution to be a good approximation to the Poisson distribution. Kass (2006) says the approximation is good for $\lambda \geq 15$. We will investigate which is correct in Homework Assignment 6. Hence consider a one second time interval and a neuron whose spike rate is on the order of 25 Hz. If its spiking activity obeys a Poisson process, then the distribution of the spiking activity on the one second interval could be approximated by the Gaussian distribution. We see already that the approximation is good for $\lambda = 40$. What are the implications of these observations for the recent paper by Wu et al. *Neural Computation* (2006) where Gaussian distributions are used to model the probability distributions of MI spiking activity?

Example 7.4 The Gaussian Approximation to the Binomial. If n is large and p is not too close to either 0 or 1, then we can approximate the cumulative distribution function for the Binomial pmf as

$$\Pr(X \le k) \approx \Phi(\frac{k-\mu}{\sigma}) \tag{7.30}$$

where $\mu = np$ and $\sigma = [np(1-p)]^{\frac{1}{2}}$. A commonly used rule of thumb that is somewhat conservative at least for $0.2 < p < 0.8$ is that it is reasonably accurate provided $np \ge 5$ and $n(1-p) \ge 5$. How does this compare to the rule for the Poisson approximation to the binomial?

Example 2.1 (**continued)**. Let us return to the Graybiel example to illustrate the effect of the CLT in the case of Bernoulli trials (which ultimately provides the Gaussian approximation to the binomial distribution). For $n = 40$ trials we have 40 values of 0 (incorrect responses) and 1 (correct responses). Suppose $p = 0.55$ then we would expect in today's experiment there would be 22 correct responses. However, when we examined the distribution of the sample mean (the proportion of correct responses) it turned out to appear very close to Gaussian. See **Figure 2D**.

We can study the central limit theorem by simulation using the binomial distribution. Let us pick $p = 0.2,$ since for $p = 0.55,$ the binomial pmf in **Figure 2D** looked very Gaussian. From **Example 2.1**, this distribution could be motivated by the observation that on the first day of the experiment, the animal had 8 correct out of 40. Therefore, we might assume that on the second day, the probability of a correct response is 0.2 for any trial.

We begin with $X_i's$ being equal to either 0 or 1. For example, with $n = 40,$ we would have 40 values of 0 or 1. Suppose that $p = 0.2,$ so that we would expect, on average, 8 of the 40 values to be 1's. Imagine a histogram of these 0's and 1's. Certainly with only two bins, these 40 values would not look Gaussian. However, when we look at the distribution of the sample mean, it turns out to look very close to Gaussian. Here, from our initial 40 0's and 1's, we would obtain one value of the sample mean, which would be the proportion of 1's in the sample.

Now we have to imagine having a large number of such samples, say 100 of them. That is, we take 100 samples of size 40 and for each sample we compute the mean of the 40 observations. If we made a histogram of those 100 sample means we would get something nicely approximating a Gaussian distribution. We can study the theoretical distribution of the sample mean. For instance, for $n = 4$ the possible values of $\bar{X}$ are 0, 0.25. 0.5, 0.75 and 1 and can be plot $p(\bar{X} = 0), p(\bar{X} = 0.25), ..., p(\bar{X} = 1)$ versus 0, 0.25. 0.5, 0.75 and 1. This is shown in the first plot in Figure 7B. This plot essentially tells us what the heights of the histogram bar would be if we did have a histogram of the 100 sample means for a sample size of $n = 4.$ The distribution of $\bar{X}$ does not look very close to Gaussian. However, as the rest of Figure 7B shows, as n increases, the distribution of $\bar{X}$ looks more and more Gaussian.

What we have just done is study the distribution of $\bar{X}$ for Bernoulli trials for several values of n with $p = 0.20.$ The distribution of $S_n = \sum_{i=1}^{n} x_i$ is Binomial and the picture of its distributions would look just like the pictures we had for the distribution of $\bar{X}$ except the x-axis would be multiplied by $n.$ In particular, as n gets large, we see the distribution looks Gaussian. This effect of the CLT may be considered as an explanation for the Gaussian approximation to the binomial.

These results have allowed us to obtain an approximate distribution for one of our most important statistics, namely the sample mean. Can we use the CLT to predict a reasonable (most probable) range for the number of correct responses we can expect to see in today's experiment if the animal does not learn? Alternatively, can we combine the CLT with the $\frac{2}{3} - 0.95$ rule to construct a predictive confidence interval for p in today's experiment? How well does the rule for assessing the accuracy of the Gaussian approximation to the binomial apply? That is, based on the rule for application of the Gaussian approximation to the binomial, would we have predicted that the Gaussian approximation would have worked well for $p = 0.55$ and $n = 40$? How good is the approximation? Can we quantify the degree of agreement or goodness-of-fit? We will answer these questions over the course of the next couple of weeks.

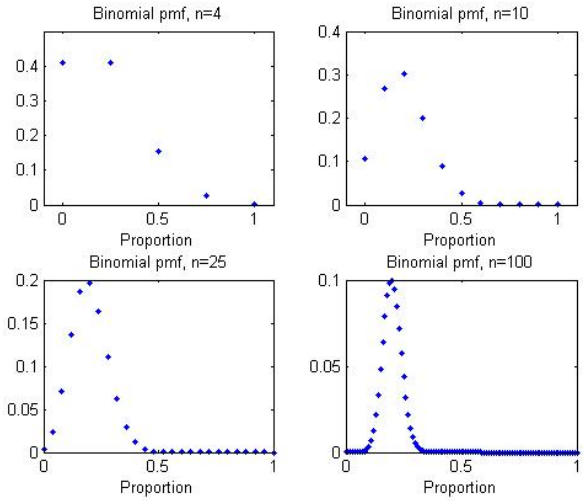

Figure 7B. Central Limit Theorem illustrated for the binomial $n = 4,10,25,100$ and $p = 0.2$.

D. Normalization in the Central Limit Theorem

The normalization for the sum of the random variables in the central limit theorem is $n^{\frac{1}{2}}$. To illustrate the importance of the choice of normalization, we consider (**Figure 7C**) sums of n random variables from a gamma distribution with $\alpha = 2$ and $\beta = 1,$ and a normalization of $n^{\frac{1}{4}}$ (column 1), of $n^{\frac{1}{2}}$ (column 2) and of n (column 3) for $n = 1,2,4,16$ and 100. When the normalization is $n^{\frac{1}{4}}$ **(Figure 7C, column 1),** we see that while the distribution is fairly symmetric by the time $n = 16$ the value of the distribution at the mode is tending to infinity. At the other extreme when the normalization is n **(Figure 7C, column 3),** the random behavior in the process is completely squashed by the time $n = 16.$ In contrast, when the normalization is $n^{\frac{1}{2}}$ **(Figure 7C, column 2)**, we see that the behavior of the process continues to become symmetric and it remains random as $n \to \infty,$ suggesting that $n^{\frac{1}{2}}$ is the correct normalization.

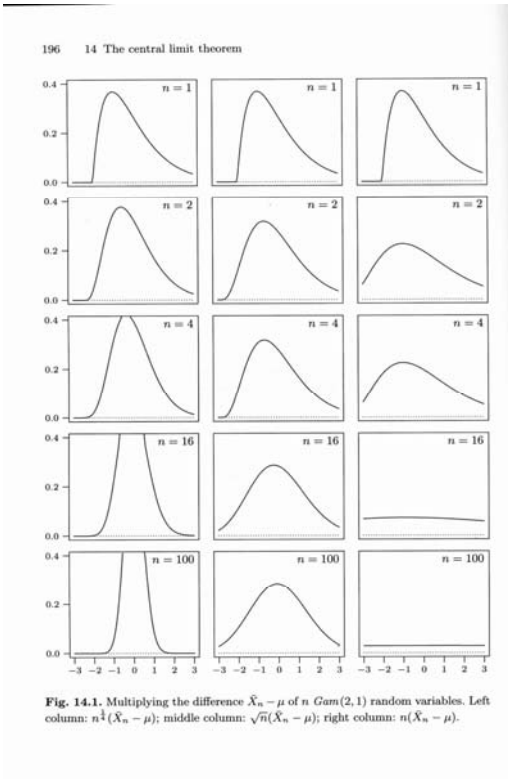

Fig. 14.1. Multiplying the difference $\bar{X}_n - \mu$ of n $Gam(2,1)$ random variables. Left column: $n^{\frac{1}{4}}(\bar{X}_n - \mu)$; middle column: $\sqrt{n}(\bar{X}_n - \mu)$; right column: $n(\bar{X}_n - \mu)$.

Figure 7C. Rate of Convergence for the average of n random draws from a gamma distribution with $\alpha = 2$ and $\beta = 1$ and a normalization of $n^{\frac{1}{4}}$ (column 1), of $n^{\frac{1}{2}}$ (column 2) and of n (column 3). Reproduced from Dekking et al. (2002).

E. Continuity Correction

Suppose that X is a discrete random variable taking on only integer values and that the pmf of X is approximated as a Gaussian random variable with mean μ and variance σ^2. Then if a and b are integers, the correction of continuity to use the Gaussian distribution is

$$\begin{aligned} \Pr(X \le b) &= \Phi([b + 1/2 - \mu]/\sigma) \\ \Pr(a \le X) &= 1 - \Phi([a - 1/2 - \mu]/\sigma) \\ \Pr(a \le X \le b) &= \Phi([b + 1/2 - \mu]/\sigma) - \Phi([a - 1/2 - \mu]/\sigma). \end{aligned} \tag{7.31}$$

instead of

$$\begin{aligned} \Pr(X \le b) &= \Phi([b - \mu]/\sigma) \\ \Pr(a \le X) &= 1 - \Phi([a - \mu]/\sigma) \\ \Pr(a \le X \le b) &= \Phi([b - \mu]/\sigma) - \Phi([a - \mu]/\sigma). \end{aligned} \tag{7.32}$$

respectively. We applied the continuity correction in the Gaussian approximation to the Poisson distribution in **Example 7.3**.

F. A More General Central Limit Theorem and an Example of Central Limit Behavior in a Physical System

We can state a more general version of the **Central Limit Theorem**.

Proposition 7.6. Central Limit Theorem (General). If $X_1, X_2, ..., X_n$ are independent random variables, possibly having different distributions but with no individual X_i making a dominant contribution to the mean $\bar{X}$, then for n sufficiently large, the distribution of $\bar{X}$ is approximately Gaussian with mean $\bar{X}$ and standard deviation $Var(\bar{X})^{\frac{1}{2}}$.

This version of the Central Limit Theorem helps to explain why the Gaussian distribution arises so often in statistical theory, and also why it seems to fit, at least roughly, so many observed phenomena. It says that whenever we average a large number of small independent effects, the result will be distributed as a Gaussian random variable. While the Central Limit Theorem involves the sample mean, it drives the large-sample behavior of most statistics: a statistic derived from a sample may usually be written, approximately, as some function (possibly a complicated function) of the sample mean, and that usually produces approximate Gaussian nature of the statistic itself. An important example is the **maximum likelihood estimator**, which is very widely used, and which we will discuss in **Lecture 9**.

Example 3.2 Magnetoencephalography (continued). Why should the MEG measurements in Figure 3E be distributed as a Gaussian random variable? It is an empirical observation that when the magnetic field inside the shielded environment is measured at a given SQUID sensor the distribution of the background measurements is Gaussian. The central limit theorem offers a theoretical reason for this observation. To see this, we note that MEG measures the magnetic fields in a given location. Anywhere there are current dipoles magnetic fields are generated. The current dipoles that we are interested in are those emanating from the brain. Remember that magnetic fields or forces are a byproduct of electric currents or moving charged particles. In any environment there are always electric currents flowing around. The SQUID sensors are exquisitely sensitive. The fields they detect are on the order of 10^{-15} Tesla (femtotesla). Therefore, if you imagine the local magnetic fields being recorded near a given SQUID sensor in a room designed to have a homogeneous magnetic field, there are a lot of local current fluctuations giving rise to local magnetic field fluctuations. If the environment is homogeneous, then no one local fluctuation will dominate (Lindeberg-Feller condition) and the local magnetic field is simply the vector sum of the local fluctuations. This later point gives the linear superposition. Hence, Gaussian structure is not surprising. Moreover, note that the observations are not independent (See Homework Assignment 6).

IV. Summary

The Law of Large Numbers states that as the number of observations in a sample of data increases the sample mean converges to the population mean whereas the Central Limit Theorem tells us that sums of random variables properly normalized can be approximated as a Gaussian distribution. This completes our section on probability theory. In the next half of the course we will study statistical theory and its application to several problems in brain and cognitive sciences making uses of the several results in probability theory we have developed.

Acknowledgments
I am grateful to Paymon Hosseini for making Figures 7A and 7C and to Uri Eden for making Figure 7B, to Julie Scott and Riccardo Barbieri for technical assistance in preparing this lecture and to Jim Mutch for very helpful proofreading and comments. Sections of this lecture have been taken with permission from the class notes for Statistical Methods in Neuroscience and Psychology written by Robert Kass in the Department of Statistics at Carnegie Mellon University. Figures 7A and 7C are reproduced from Dekking et al. (2002).

References

DeGroot MH, Schervish MJ. *Probability and Statistics*, 3rd edition. Boston, MA: Addison Wesley, 2002.

Dekking FM, Kraaikamp C, Lopuhaa HP, Meester LE. *A Modern Introduction to Probability and Statistics*. London: Springer-Verlag, 2002.

Pawitan Y. *In All Likelihood: Statistical Modeling and Inference Using Likelihood.* London: Oxford, 2001.

Port SC. *Theoretical Probability for Applications.* New York: John Wiley & Sons, 1994.

Rice JA. *Mathematical Statistics and Data Analysis*, 3rd edition. Boston, MA, 2007.

Literature Reference

Wu W, Gao Y, Bienenstock E, Donoghue JP, Black MJ. Bayesian population decoding of motor cortical activity using a Kalman filter. *Neural Computation* 18(1):80-118, 2006.

Website Reference

Kass RE, Chapter 2: Probability and Random Variables
http://lib.stat.cmu.edu/~kass/smn/notes/2.pdf