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

Lecture 8 Estimation Theory: Method-of-Moments Estimation

I. Objectives

Understand the statistical paradigm for reasoning under uncertainty.

Understand the principle of estimation in the statistical paradigm.

Learn how to construct method-of-moments estimates for standard probability models.

Understand how we assess the uncertainty in the method-of-moments estimates in terms of confidence intervals.

Having completed our section on probability theory, we will study statistics in the balance of the course. We review the logic of the statistics paradigm.

II. The Statistics Paradigm

Statistics is the science of making decisions under uncertainty using mathematical models derived from probability theory.

Given a sample of data, $x_1,...,x_n$, a **statistic** is any function of the data. For example the mean, median, mode, standard deviation, variance, kurtosis, and the skewness are all statistics.

A. THE STATISTICAL PARADIGM (Box, Tukey)

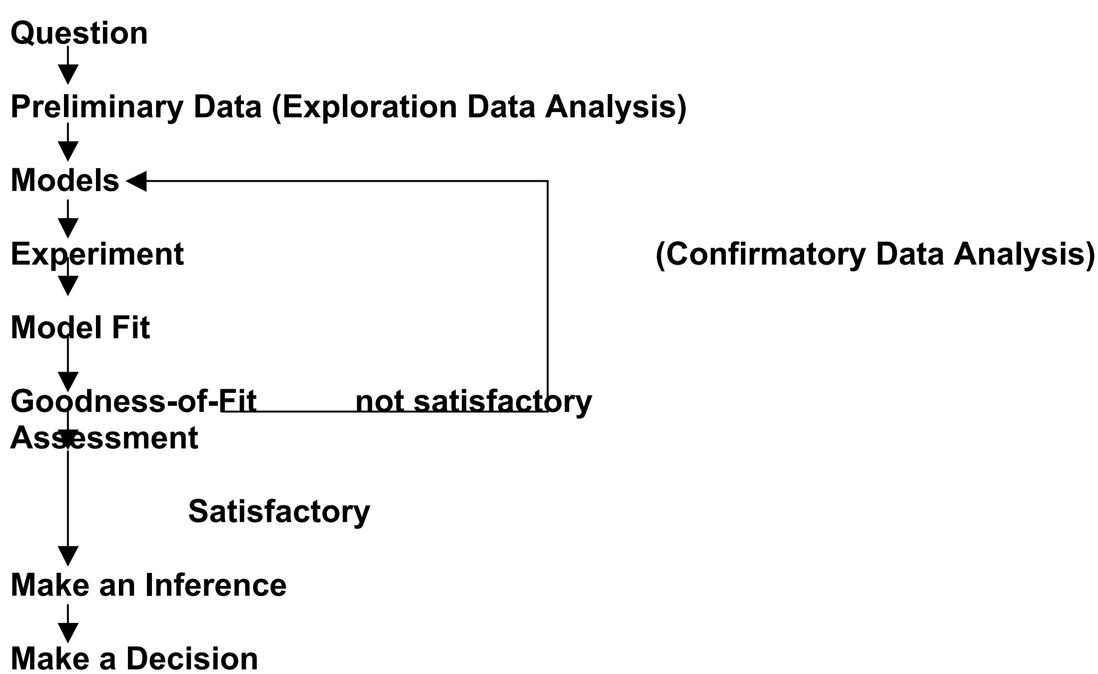

Figure 8A. The Box-Tukey paradigm for iterative model building based on the concepts of exploratory and confirmatory data analysis.

The **Box-Tukey Paradigm** (Figure 8A) defines an 8-step approach to experimental design statistical model building and data analysis. The iterative cycle of this approach to statistical model building, data analysis and inference is attributed to George Box.

The initial steps are termed **exploratory data analysis** (EDA). The idea of EDA was formalized by John Tukey. In this phase the statistical model builder uses whatever knowledge is available to propose an initial model or set of models for the data to be analyzed. In this phase a set of descriptive, usually non-model based statistical tools, are used to perform a preliminary analysis of the data. This may entail analysis of data from the literature or preliminary data collected in the current laboratory. The objective of the EDA phase of the analysis is to develop a set of initial probability models that may be used to describe the stochastic (random) structure in the data. The EDA techniques we have learned thus far are data plots, histograms, five-number summaries, stem-and-leaf plots and box plots.

Confirmatory data analysis (CDA) uses formal statistical model fitting and assessments of goodness-of-fit to evaluate how well a model fits the data. Once a satisfactory fit to the data is obtained as evaluated in a goodness-of-fit assessment, the scientist may go ahead to make an inference and then a decision. The first part of CDA is proposing an appropriate probability model which is why we just learned a large set of probability models and their properties. We now learn how to use these probability models and extensions of them to perform statistical analyses of data. We will develop our ability to execute the steps in this paradigm and hence, use it to conduct statistical analyses during the next seven weeks.

III. Method-of-Moments Estimation

A. What is Estimation?

Once a probability model is specified for a given problem, the next step is estimation of the model parameters. By **estimation** we mean using a formal procedure to compute the model parameter from the observed data. We term any formal procedure that tells us how to compute a model parameter from a sample of data an **estimator**. We term the value computed from the application of the procedure to actual data the **estimate**.

In the next series lectures we will study three methods of estimation: method-of-moments, maximum likelihood and Bayesian estimation.

Example 2.1 (continued). On the previous day of the learning experiments the animal executed 22 of the 40 trials correctly. A reasonable probability model for this problem is the binomial distribution with $n = 40$ and p unknown, if we assume the set of trials to be independent. We simply took $\hat{p}$, the estimate of p to be $22/40 = 0.55.$ While this seems reasonable, in what sense is this the "best" estimate? Is this still the best estimate if the trials are not independent?

Example 3.2 (continued). The MEG measurements appeared Gaussian in shape (**Figure 3E**). This was further supported by the $Q-Q$ plot goodness-of-fit analysis in **Figure 3H**. The Gaussian distribution has two parameters μ and σ^2. We estimated μ as $\bar{x} = -0.2 \cdot 10^{-11}$ and $\hat{\sigma}^2 = 2.0 \cdot 10^{-22}$, where

$$\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

where we recall that $\bar{x}$ and $\hat{\sigma}^2$ are the sample mean and sample variance respectively.

Again, are these the best estimates of the parameters? The plot of the estimated probability density in Figure 3E and the $Q-Q$ plot in Figure 3H suggest that these estimates are very reasonable, i.e. highly consistent with the data.

Example 8.1. Spiking Activity of Retinal Neurons (Iyengar and Liu, 1997). Retinal neurons are grown in culture under constant light and environmental conditions. The spontaneous spiking activity of the neurons is recorded. The investigators wanted to develop a statistical model that accurately describes the stochastic structure of this activity. That is, the interspike interval distribution. The gamma and inverse Gaussian distributions have been suggested as possible probability models. How do we compute the parameters of these probability models from the observed measurements? We perform first an exploratory data analysis by making the boxplot, the five-number summary and to aid intuition with the interpretation of the boxplot (**Figure 8B**), we include a histogram (**Figure 8C**).

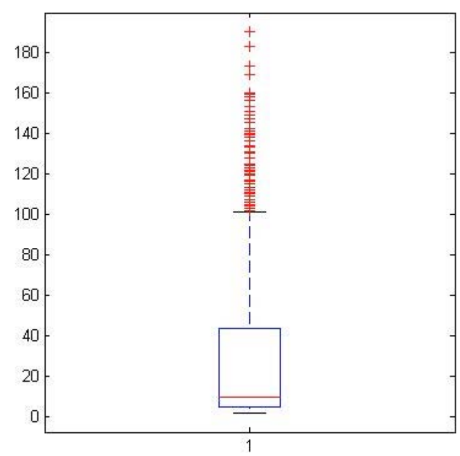

Figure 8B. Boxplot of Interspike Interval of Spiking Activity of Retinal Neurons under Constant Conditions in Slice (Iyengar and Liu, Biol. Cyber. 1997).

The five-number summary for the retinal neuron spiking data is: the minimum is 2 msec, the 25th percentile is 5 msec, the median is 10 msec, the 75th percentile is 43 msec, and the maximum is 190 msec.

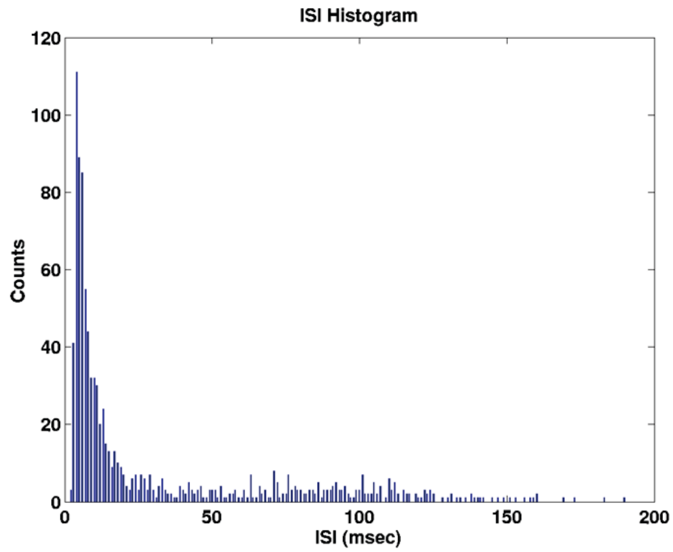

Figure 8C. Interspike Interval Histogram of Spiking Activity of Retinal Neurons under Constant Conditions in Slice (Iyengar and Liu, Biol. Cyber. 1997).

For each of these examples, how do we measure the uncertainty in these parameter estimates?

B. Method-of-Moments Estimation

One very straight forward, intuitively appealing approach to estimation is the **method-of -moments**. It is simple to apply. Recall that the theoretical moments were defined in **Lecture 6** as follows

Given $p(x)$ a probability mass function we define the i^{th} **theoretical moment** as

$$\mu_i = \sum_j x_j^i p(x_j), \tag{8.1}$$

for $i = 1, 2, \ldots,$ and given $f(x)$ a probability density function, we define the i^{th} **theoretical moment** as

$$\mu_i = \int x^i f(x) dx, \tag{8.2}$$

for $i = 1, 2, \ldots,$. These are the moments about zero, or the non-central moments. Note that the first moment is simply the mean and that the variance is

$$\begin{aligned}\sigma^2 &= E(x-\mu)^2 \\ &= \int (x-\mu)^2 f(x)dx \\ &= \int (x^2 - 2\mu x + \mu^2) f(x)dx \\ &= \int x^2 f(x)dx - \mu^2 \int f(x)dx \\ &= \mu_2 - \mu_1^2.\end{aligned} \tag{8.3}$$

The same result holds for discrete random variables with the integrals changed to summations. The variance is the second central moment. It is imperative that we add at this point that these definitions only make sense if these moments exist. There are certainly cases of probability densities that are well defined for which the moments do not exist. One often quoted example is the Cauchy distribution defined as

$$f(x) = \frac{1}{2\pi(1+x^2)}, \tag{8.4}$$

for $x \in (-\infty, \infty)$.

Given a sample of data $x_1,...,x_n$, the i^{th} **sample moment** is

$$\hat{\mu}_i = n^{-1} \sum_{r=1}^{n} x_r^i \tag{8.5}$$

for $i = 1,2,3,...,$. The sample moments are computed by placing mass of n^{-1} at each data point. Because all the numbers are finite and the sum is finite, the sample moment of each order i exists.

Method-of-Moments Estimation. Assume we observe $x_1,...,x_n$, as a sample from $f(x|\theta)$, a probability model with a $d-$dimensional unknown parameter θ. Assume that the first d moments of $f(x|\theta)$ are finite. Equate the theoretical moments to the corresponding sample moments and estimate θ as the solution to the $d-$dimensional equation system

$$\mu_i(\theta) = \hat{\mu}_i \tag{8.6}$$

for $i = 1,...,d$. The estimate of θ computed from Eq. 8.6 is the **method-of-moments estimate** and we denote it as $\hat{\theta}_{MM}$.

Example 8.1 (continued). To fit the gamma probability model to the retinal ganglion cell interspike model using the method-of-moments, we consider

$$\mu = E(X) = \alpha\beta^{-1} \tag{8.7}$$

$$\sigma^2 = E(X-\mu)^2 = \alpha\beta^{-2} \tag{8.8}$$

Note that we have used the second central moment instead of the second moment about zero to simplify the computation. Because $\theta = (\alpha, \beta)$ is 2-dimensional, we compute the first sample moment and the second sample central moment as

$$\bar{x} = n^{-1} \sum_{r=1}^{n} x_r \tag{8.9}$$

$$\hat{\sigma}^2 = n^{-1} \sum_{r=1}^{n} (x_r - \bar{x})^2 \tag{8.10}$$

and our 2-dimensional system of equations is

$$\bar{x} = \alpha \beta^{-1} \tag{8.11}$$

$$\hat{\sigma}^2 = \alpha \beta^{-2} \tag{8.12}$$

Solving Eqs. 8.11 and 8.12 yields the method-of-moments estimates

$$\begin{aligned} \hat{\beta}_{MM} &= \bar{x}(\hat{\sigma}^2)^{-1} \\ \hat{\alpha}_{MM} &= \bar{x}^2(\hat{\sigma}^2)^{-1}. \end{aligned} \tag{8.13}$$

Example 2.1 (continued). To fit the binomial probability model to the rat learning data using the method-of-moments we have $\bar{x} = n^{-1} \sum_{r=1}^{n} x_i$ and

$$\bar{x} = p \tag{8.14}$$

The one-dimensional equation is

$$\hat{p}_{MM} = \bar{x} \tag{8.15}$$

and

$$\hat{p}_{MM} = n^{-1} \sum_{r=1}^{n} x_r = (40)^{-1} 22 = 0.55. \tag{8.16}$$

Example 3.2 (continued). To fit the Gaussian probability model to the MEG noise data using method-of-moments we have simply

$$\begin{aligned} \hat{\mu}_{MM} &= \bar{x} \\ \hat{\sigma}^2_{MM} &= n^{-1} \sum_{r=1}^{n} (x_i - \hat{\mu}_{MM})^2 \end{aligned} \tag{8.17}$$

Remark 8.1. The method-of-moments estimates are simple to compute. As we showed, they may not exist. They are not unique. To see this latter point, note that we have taken the first d

moments to construct the estimation equations but in principle and in practice, we could use any d theoretical and d sample moments. To see this consider $x_1,...,x_n$ as a sample from a Poisson distribution with unknown parameter λ. Recall that for a Poisson distribution

$$E(X) = \lambda \tag{8.18}$$

and

$$Var(X) = \lambda. \tag{8.19}$$

Hence, we have

$$\lambda_{MM} = \hat{\mu}_1 = n^{-1}\sum_{r=1}^{n} x_r \tag{8.20}$$

$$\lambda_{MM} = \hat{\mu}_2 - (\hat{\mu}_1)^2 = n^{-1}\sum_{r=1}^{n} x_r^2 - (n^{-1}\sum_{r=1}^{n} x_r)^2. \tag{8.21}$$

Is one of these estimates "better"? We will discuss this further when we discuss maximum likelihood estimation (See Kass, Ventura and Brown, 2005).

C. Uncertainty in the Method-of-Moments Estimate
1. Confidence Interval for the Mean

We computed above the estimate of the probability of a correct response as $\hat{p} = 0.55$ for the learning experiment in **Example 2.1**. The method-of-moments procedure requires a justification. We certainly do not believe that this is the exact answer. There is obvious uncertainty and we would like to quantify that uncertainty. This leads us naturally to a discussion of the error in this estimate or alternatively a confidence interval for the true parameter.

We recall from the Central Limit Theorem (**Lecture 7**) that if $X_1,...,X_n$ is a sample from a distribution with mean μ and variance σ^2 then for n large

$$\bar{X} : N(\mu, \frac{\sigma^2}{n}) \tag{8.22}$$

We can use this fact to construct a confidence interval for μ as a way of measuring the uncertainty in $\bar{X}$, which for the Example 2.1 is $\hat{p}_{MM}$. Assume that σ^2 is known. Then from the Central Limit Theorem, if we take $Z = \frac{n^{\frac{1}{2}}(\bar{X} - \mu)}{\sigma}$ then we have

$$\begin{aligned}
\Pr(|Z| < 1.96) = 0.95 & \\
&= \Pr(-1.96 \le Z \le 1.96) \\
&= \Pr\left(-1.96 \le \frac{n^{\frac{1}{2}}(\bar{X}-\mu)}{\sigma} \le 1.96\right) \\
&= \Pr\left(\frac{-1.96\sigma}{n^{\frac{1}{2}}} \le \bar{X}-\mu \le \frac{1.96\sigma}{n^{\frac{1}{2}}}\right) \\
&= \Pr\left(\frac{1.96\sigma}{n^{\frac{1}{2}}} \ge \mu-\bar{X} \ge \frac{-1.96\sigma}{n^{\frac{1}{2}}}\right) \\
&= \Pr\left(\bar{X}-\frac{1.96\sigma}{n^{\frac{1}{2}}} \le \mu \le \bar{X}+\frac{1.96\sigma}{n^{\frac{1}{2}}}\right)
\end{aligned} \tag{8.23}$$

Since $1.96 \approx 2$, we have the frequently quoted statement that a 95% confidence interval for the mean of a Gaussian distribution is

$$\bar{X}-\frac{2\sigma}{n^{\frac{1}{2}}} \le \mu \le \bar{X}+\frac{2\sigma}{n^{\frac{1}{2}}}. \tag{8.24}$$

The quantity $\frac{\sigma}{n^{\frac{1}{2}}}$ is the standard error (SE) of the mean. This is simply an application of Kass' 2/3-95% rule applied to the Gaussian approximation from the Central Limit Theorem of the distribution of $\bar{X}$ (**Lecture 3**).

Remark 8.2. If the random sample is from a Gaussian distribution, then the result is exact.

Remark 8.3. If the random sample is from a binomial distribution (**Example 2.1**), then we have $\hat{p}=\bar{x}$ and $Var(\hat{p}) = Var(\bar{x}) = \frac{\sigma^2}{n} = \frac{p(1-p)}{n}$. Hence, the approximate 95% confidence interval is

$$\hat{p} \pm 2\left[\frac{\hat{p}(1-\hat{p})}{n}\right]^{\frac{1}{2}} \tag{8.25}$$

For the learning experiment in **Example 2.1**, this becomes

$$\begin{aligned}
&0.55 \pm 2\left[\frac{0.55(.45)}{40}\right]^{\frac{1}{2}} \\
&0.55 \pm 2 \times 0.078
\end{aligned} \tag{8.26}$$

Hence, the confidence interval for p is [0.394, 0.706]. Also notice that $p(1-p) \le \frac{1}{4}$ for all p so that an often used upper bound interval is

$$\hat{p} \pm 2\left[\frac{1}{4n}\right]^{\frac{1}{2}} \tag{8.27}$$

or

$$\hat{p} \pm \frac{1}{n^{\frac{1}{2}}}. \tag{8.28}$$

Example 8.2. Choosing n to set the length of the confidence interval. How many samples should be taken to insure that the length of the 95% confidence interval is less than 0.1? The length of the confidence interval is $4[\frac{p(1-p)}{n}]^{\frac{1}{2}} \le 4[\frac{1}{4n}]^{\frac{1}{2}} = 2[\frac{1}{n}]^{\frac{1}{2}}$. If we set $2[\frac{1}{n}]^{\frac{1}{2}}$ = 0.1, Then we find that $n = 400$. Notice that this choice of sample size is independent of the true value of p.

Remark 8.4. If the random sample is from a Poisson distribution then we have that the confidence interval is

$$\hat{\lambda} \pm 2\left(\frac{\hat{\lambda}}{n}\right)^{\frac{1}{2}}. \tag{8.29}$$

What are the conditions for this result to be reasonable?

Remark 8.5. We can construct 90% or 99% confidence intervals, or in general intervals of any desired percentage. The chosen percentage is the **confidence level.** This is often confused with the significance level of a statistical test. These concepts are related but different. We will discuss this in **Lecture 12**. For a 90% confidence interval we have

$$\bar{x} \pm \frac{1.64\sigma}{n^{\frac{1}{2}}}, \tag{8.30}$$

and for a 99% confidence interval we have

$$\bar{x} \pm \frac{2.57\sigma}{n^{\frac{1}{2}}}. \tag{8.31}$$

For Gaussian and symmetric distributions we construct the confidence intervals symmetrically. In general this need not be the case.

2. The Classical Interpretation of the Confidence Interval

The classical (frequentist) interpretation of the confidence interval is that stated in terms of the operating characteristics of the confidence interval. That is, if we were to estimate μ repeatedly with new data then, in the long-run (with infinitely many repetitions), about 95% of the time the interval $\bar{x} \pm \frac{2\sigma}{n^{\frac{1}{2}}}$ would cover μ. In other words if we considered using the procedure again and again, it would work in the sense that the confidence we state would correspond, more or less, to the actual probability that the confidence interval would cover the unknown values of μ.

We can illustrate the long-run behavior in a simulation study. Let us return to the Example 2.1 and assume that $n = 40$ and suppose that the true $p = 0.6.$ That is, the animal's true propensity to respond correctly is 0.6. It is easy to simulate 10,000 Binomial observations from the Binomial distribution with $n = 40$ and $p = 0.6.$ We can ask, how often does the interval in Eq. 8.25 contain the true value of $0.6.$ The simulation takes the following steps:

1. For i =1 to 10,000
   Generate $X_i \sim Binomial(n=40, p=0.6)$
   Compute $\hat{p}_i = X_i / 40$
   Compute $se_i = (\hat{p}_i(1-\hat{p}_i)/40)^{\frac{1}{2}}$
2. Sum the number of i for which $\hat{p}_i - 2se_i < 0.6 < \hat{p}_i + 2se_i$.

We did this and obtained 9,454 as the sum in step 2. As a result, we conclude that to two-digit accuracy, the true value of p is covered by the interval in Eq. 8.25 95% of the time. We therefore see the sense in which the standard procedure is well calibrated for large samples.

One way probability is often interpreted is in terms of bets: if we say an event has probability of 0.5, this indicates a willingness to take either side of a bet that the event will happen. If it has probability 0.95, this indicates a willingness to take either side of a bet that gives 19:1 odds the event will happen. Here, if we were to take either side of a bet that gives 19:1 odds the that the interval will cover the true value of p, in the long run (playing this game a very large number of times) we would expect to neither win nor lose money but rather to more-or-less break even. In this sense, confidence can be interpreted as making a statement about the operating characteristics of the confidence interval procedure. We will review other interpretations of the intervals.

3. Confidence Interval for the Variance of a Gaussian Distribution

These results have allowed us to compute the approximate distribution of the mean of a random sample from a distribution with finite mean and variance. What is the distribution of the sample variance? While we will not answer this question generally, we will answer it for the case of a Gaussian distribution. If $x_1, x_2, ..., x_n$ is a sample from a Gaussian distribution with mean μ and variance σ^2, and we define $\hat{\sigma}^2 = n^{-1}\sum_{r=1}^{n}(x_r - \bar{x})^2$ then

$$\frac{n\hat{\sigma}^2}{\sigma^2} = \sum_{r=1}^{n}\frac{(x_r - \bar{x})^2}{\sigma^2} \sim \chi^2_{(n-1)}. \qquad (8.32)$$

We establish the result in the **Addendum to Lecture 8**. We can use this fact to construct a confidence interval

$$\begin{aligned}\Pr(\chi^2_{n-1}(0.025) \le \tfrac{n\hat{\sigma}^2}{\sigma^2} \le \chi^2_{n-1}(0.975)) &= \Pr(\frac{1}{\chi^2_{n-1}(0.025)} \ge \tfrac{\sigma^2}{n\hat{\sigma}^2} \ge \frac{1}{\chi^2_{n-1}(0.975)}) \\ &= \Pr(\frac{n\hat{\sigma}^2}{\chi^2_{n-1}(0.975)} \le \sigma^2 \le \frac{n\hat{\sigma}^2}{\chi^2_{n-1}(0.025)})\end{aligned} \qquad (8.33)$$

Remember that the χ^2 distribution with $n-1$ degrees of freedom is a gamma distribution with $\alpha = \frac{(n-1)}{2}$ and $\beta = \frac{1}{2}$.

D. The Gaussian Distribution Revisited

When we do not know the value of σ (as is almost always the case), we can estimate it with the sample standard deviation, usually denoted by $\hat{\sigma}$, or $s = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ and use either of these in place of σ. This results in an approximate confidence interval which applies for n large. The approximate standard error of the mean $\bar{x}$ becomes

$$se(\bar{x}) = \frac{\hat{\sigma}}{n^{\frac{1}{2}}}. \tag{8.34}$$

The approximate 95% confidence interval is

$$\bar{X} - \frac{1.96\hat{\sigma}}{n^{\frac{1}{2}}} \le \mu \le \bar{X} + \frac{1.96\hat{\sigma}}{n^{\frac{1}{2}}} \tag{8.35}$$

The approximation improves when the data are more nearly Gaussian and may not be as good in moderate sample sizes (say, 25 observations) with strongly non-Gaussian data. Hence, it is important to check the Gaussian assumption with $Q-Q$ plots.

For small samples and Gaussian data, it is possible to adjust the confidence intervals for the estimation of σ by s. The interval becomes slightly wider than the "exact" interval in Eq. 8.23. We are less certain about μ when there is added uncertainty about σ^2. The correction uses the $t-$distribution. The $t-$distribution is a unimodal symmetric distribution which looks very much like the Gaussian. The particular $t-$distribution used depends on the sample size used to compute $\bar{x}$ and s^2. The distributions are lower in the middle and have thicker tails than the Gaussian. If the sample size is n then $\nu = n-1$ is the number of degrees of freedom, and the appropriate $t-$distribution to use is the $t-$distribution with ν degrees of freedom. Because the tails of the $t-$distribution are fatter, the value from the $t-$distribution corresponding to 1.96 on the Gaussian distribution will be larger (See Table 8.1 and Figure 8D). The 95% confidence interval based on the $t-$distribution if $x_1, x_2, ..., x_n$ is a sample from a Gaussian distribution with mean μ and variance σ^2 is

$$\bar{x} - t_{0.975,n-1}\frac{s}{n^{\frac{1}{2}}} \le \mu \le \bar{x} + t_{0.975,n-1}\frac{s}{n^{\frac{1}{2}}} \tag{8.36}$$

As n increases $t_{0.975}$ goes to 1.96. In general as n increases the $t-$distribution on ν degrees of freedom converges to a Gaussian distribution with mean 0 and variance 1.

ν	0.975
10	2.228
15	2.131
20	2.086
25	2.060
30	2.042
40	2.021
60	2.000
120	1.980
∞	1.960

Table 8.1. The 0.975 quantiles of the t-distributions as a function of degrees of freedom.

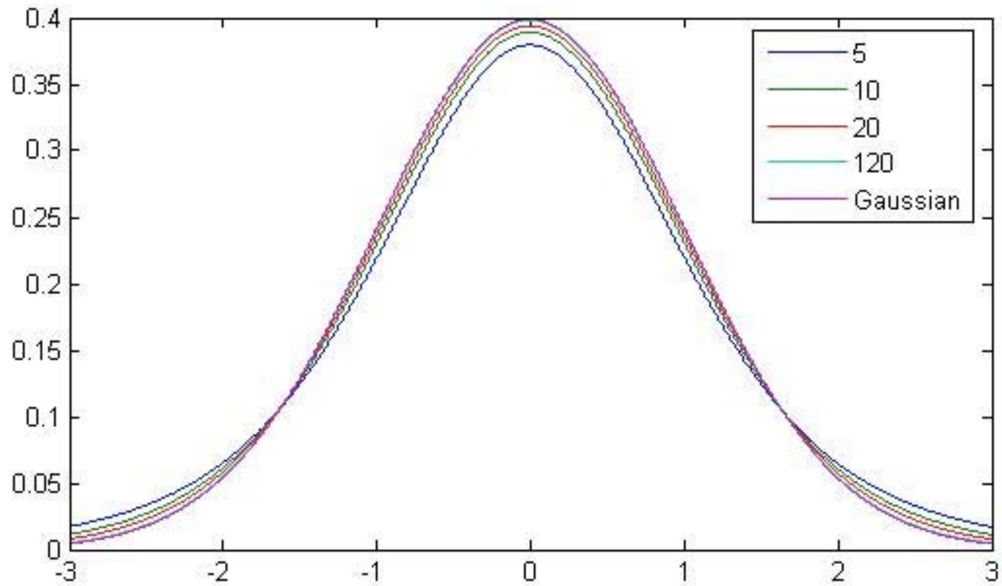

Figure 8D. Family of t-distributions plotted as a function of the number of degrees of freedom ν. The limiting distribution is the standard Gaussian.

For sample sizes of 15 and less the t correction is important.

IV. Summary

Method-of-moments is a simple intuitive approach to estimating the parameters in a probability density. We will analyze further the extent to which these methods are "best" once we have studied likelihood theory (**Lecture 9**) and Bayesian methods (**Lecture 10**). We will use the bootstrap (**Lecture 11**) to compute estimates of uncertainty for the method-of-moments estimates.

Acknowledgments

I am grateful to Uri Eden for making the figures, to Jim Mutch for careful proofreading and comments and to Julie Scott for technical assistance. Several sections of this lecture have been taken with permission from the class notes for the course Statistical Methods in Neuroscience and Psychology written by Robert Kass in the Department of Statistics at Carnegie Mellon University.

Textbook Reference

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

Iyengar S, Liao Q. Modeling neural activity using the generalized inverse Gaussian distribution. *Biological Cybernetics*, 1997, 77: 289–295.

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

Literature Reference

Kass RE, Ventura V, Brown EN. Statistical issues in the analysis of neuronal data. *Journal of Neurophysiology*, 2005, 94: 8-25.

Website Reference

Kass RE, Chapter 3: Estimation, 2006.

9.07 Statistics for Brain and Cognitive Science
Fall 2016