Lectures on EE 3025: Statistics

John Kieffer
Department of Electrical and Computer Engineering
University of Minnesota Twin Cities
Minneapolis, MN 55455

Lecture 26

Statistics Part 1

We are now entering the "statistics" part of EE 3025. During the next few lectures, we will examine some selected topics from Chapters 6,7, and 9 having to do with statistics. As we indicated in our "Preview" at the end of Lecture 25, our first task will be coverage of the CLT, LLN, and confidence interval design.

26.1 Examples of Statistics

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from a sampling distribution which has mean μ and variance σ^2. A *statistic* is any RV which is a function of this random sample. When you perform your experiment, the observed value for a statistic must be computable from the observed values of $X_1, X_2, \cdots, X_n$; in particular, a statistic cannot depend on any unknown parameters.

Examples of Statistics

- The sum $X_1 + X_2 + \cdots + X_n$ of the random sample values is a statistic called S_n:

$$S_n \triangleq X_1 + X_2 + X_3 + \cdots + X_n.$$

- The average of the X_i's is a statistic called the *sample mean*. It is denoted by $\bar{X}_n$ (when we want to make clear that the sample size is n) or $\bar{X}$ (when the sample size is clear). It is defined by:

$$\bar{X}_n = \bar{X} \triangleq \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

- The *sample variance* is a statistic. It is defined by

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}. \tag{26.1}$$

For large n, this is roughly the same thing as

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

We will explain later why statisticians like to divide by $n-1$ instead of n when defining the sample variance.

- If the mean μ is known, then

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n} \tag{26.2}$$

is a legitimate statistic.

For the present, our focus is on the statistics S_n and $\bar{X}_n$. The statistic S_n is important because of its presence in the statement of the central limit theorem. The sample mean statistic $\bar{X}_n$ is important because it can be used to estimate the mean μ when μ is unknown.

A bit later, I will talk about the statistics (26.1) and (26.2), which are used to estimate the variance σ^2.

26.2 Mean and Variance of S_n and $\bar{X}_n$

Useful Result

(a): The statistic S_n has mean and variance as follows:

$$E[S_n] = n\mu, \quad Var(S_n) = n\sigma^2.$$

(b): The sample mean $\bar{X}_n$ has mean and variance as follows:

$$E[\bar{X}_n] = \mu, \quad Var(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Proof of (a). You can always take the expected value of a sum term by term:

$$E[S_n] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \mu = n\mu.$$

In general, you cannot take the variance of a sum term by term. However, if the terms in the sum are independent, then we know that you can take the variance term by term. The terms X_i in the random sample $X_1, X_2, \cdots, X_n$ are independent by definition. Therefore,

$$Var[S_n] = Var[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} Var[X_i] = \sum_{i=1}^{n} \sigma^2 = n\sigma^2.$$

Proof of (b). We can easily derive the mean and variance of the sample mean $\bar{X}_n$ from the mean and variance of S_n, because the sample mean is a scalar multiple of S_n, and we know what happens to mean and variance when we take a scalar multiple:

$$\begin{aligned} E[\bar{X}_n] &= E[S_n/n] = (1/n)E[S_n] = (1/n)(n\mu) = \mu. \\ Var[\bar{X}_n] &= Var[S_n/n] = (1/n)^2 Var(S_n) = (1/n^2)(n\sigma^2) = \sigma^2/n \end{aligned}$$

(Recall that when you pull a scalar out of a variance operator, you have to square the scalar.)

Remarks

(a): If a RV Y has mean μ_Y and variance σ_Y^2, recall from Chapters 2-3 that the RV

$$\frac{Y - \mu_Y}{\sigma_Y}$$

has mean 0 and variance 1. If we set $Y = S_n$, then we conclude that the "normalized sum"

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \tag{26.3}$$

has mean 0 and variance 1. If we set $Y = \bar{X}_n$, then we conclude that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \tag{26.4}$$

has mean 0 and variance 1.

(b): Actually, the two expressions (26.3) and (26.4) are equal to one another:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{(S_n - n\mu)/n}{(\sigma\sqrt{n})/n} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}.$$

(c): By definition of variance,

$$E[(\bar{X}_n - E[\bar{X}_n])^2] = Var(\bar{X}_n).$$

Plugging in what the mean and variance of $\bar{X}_n$ are, we conclude that

$$E[(\bar{X}_n - \mu)^2] = \frac{\sigma^2}{n}. \tag{26.5}$$

Equation (26.5) is important for the following reason: Suppose we want to use the sample mean $\bar{X}_n$ to estimate μ when μ is unknown. As we select random samples from our sampling distribution on trial after trial and compute the different sample mean values, we will see that

$\bar{X}_n$ will fluctuate on either side of μ, which is in a fixed position on the real line. The quantity $E[(\bar{X}_n - \mu)^2]$ quantifies how big these fluctuations can be, on average, in a mean-square sense. (You can average up the squares of the differences between the observed values of $\bar{X}_n$ and μ over a large number of trials; this will be approximately what $E[(\bar{X}_n - \mu)^2]$ is.) Notice from (26.5) that $E[(\bar{X}_n - \mu)^2]$ is getting smaller and smaller as the sample size n increases; since σ^2/n approaches zero as n goes to infinity, we can fix a sample size n so large that $E[(\bar{X}_n - \mu)^2]$ will be smaller than whatever preset positive quantity you want. In this way, you see that formula (26.5) tells us that the sample mean $\bar{X}_n$ becomes a better and better estimator of μ the larger we take the sample size n.

26.3 Probabilistic Behavior of S_n, $\bar{X}_n$: the CLT

We want to know what we can say about the probability distribution of the quantities

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}. \tag{26.6}$$

First, we investigate this question for a Gaussian sampling distribution and then we investigate this question for a nonGaussian sampling distribution.

26.3.1 Case of Gaussian sampling distribution

In this case, we know that S_n is Gaussian. (We proved earlier using moment generating function techniques that the sum of independent Gaussian RV's is also Gaussian.) If you translate and/or scale a Gaussian RV, you get another Gaussian RV. (We know this from the Chapter 2-3 material.)

We immediately conclude that the quantities (26.6) are both Gaussian$(0, 1)$ RV's (i.e., standard Gaussian RV's). Therefore, we can write

$$\begin{aligned} P\left[a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b\right] &= \int_a^b \left(\frac{1}{\sqrt{2\pi}}\right) \exp(-z^2/2)dz \\ P\left[a \le \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \le b\right] &= \int_a^b \left(\frac{1}{\sqrt{2\pi}}\right) \exp(-z^2/2)dz \end{aligned}$$

These equations are true for every sample size n. Doing some algebra on the left side of the second equation, you can re-write the second equation as

$$P\left[\mu + \frac{a\sigma}{\sqrt{n}} \le \bar{X}_n \le \mu + \frac{b\sigma}{\sqrt{n}}\right] = \int_a^b \left(\frac{1}{\sqrt{2\pi}}\right) \exp(-z^2/2)dz$$

26.3.2 Case of NonGaussian sampling distribution

We are now sampling from a nonGaussian distribution. In this case, the central limit theorem (CLT) can be applied. (See Theorem 6.14 on page 258 for a staterment of the CLT.) Roughly speaking, the CLT tells us that the random quantities (26.6) have approximately a Gaussian$(0,1)$ distribution if n is large:

$$\begin{aligned} \frac{S_n - n\mu}{\sigma\sqrt{n}} &\approx Gaussian(0,1), \quad n\, large \\ \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} &\approx Gaussian(0,1), \quad n\, large \end{aligned}$$

This allows us to say that for large n, the probability statements given in Section 26.3.1 are approximately true, that is:

$$\begin{aligned} P\left[a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b\right] &\approx \int_a^b \left(\frac{1}{\sqrt{2\pi}}\right) \exp(-z^2/2)dz \\ P\left[a \le \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \le b\right] &\approx \int_a^b \left(\frac{1}{\sqrt{2\pi}}\right) \exp(-z^2/2)dz \\ P\left[\mu + \frac{a\sigma}{\sqrt{n}} \le \bar{X}_n \le \mu + \frac{b\sigma}{\sqrt{n}}\right] &\approx \int_a^b \left(\frac{1}{\sqrt{2\pi}}\right) \exp(-z^2/2)dz \end{aligned}$$

More precisely, as $n \to \infty$, the probabilities on the left become Gaussian probabilities in the limit:

$$\begin{aligned} \lim_{n\to\infty} P\left[a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b\right] &= \int_a^b \left(\frac{1}{\sqrt{2\pi}}\right) \exp(-z^2/2)dz \\ \lim_{n\to\infty} P\left[a \le \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \le b\right] &= \int_a^b \left(\frac{1}{\sqrt{2\pi}}\right) \exp(-z^2/2)dz \\ \lim_{n\to\infty} P\left[\mu + \frac{a\sigma}{\sqrt{n}} \le \bar{X}_n \le \mu + \frac{b\sigma}{\sqrt{n}}\right] &= \int_a^b \left(\frac{1}{\sqrt{2\pi}}\right) \exp(-z^2/2)dz \end{aligned}$$

A completely general proof of the CLT (which would work for every single possible sampling distribution) is somewhat tricky. Instead, in a later lecture, I will prove a special case of the CLT for an easily handled type of sampling distribution. You should (hopefully) find this proof partially convincing concerning why the CLT is true. Also, the reader can refer to Recitation 9 Matlab demos illustrating the CLT for various sampling distributions.

26.4 LLN as special case of CLT

The law of large numbers (LLN) says that, regardless of the sampling distribution, the following limiting relation is true:

$$\lim_{n\to\infty} P[\mu - \epsilon \le \bar{X}_n \le \mu + \epsilon] = 1, \quad for\ every\ \epsilon > 0. \tag{26.7}$$

The purpose of this section is to show you that statement (26.7) is true using the CLT.

First, let us discuss a little bit the type of converence exhibited in statement (26.7). Statisticians call this type of convergence *stochastic convergence*. More generally, if we have an infinite sequence of RV's

$$Y_1, Y_2, Y_3, \cdots,$$

then we say that this sequence *converges stochastically* to a parameter θ if

$$\lim_{n\to\infty} P[\theta - \epsilon \leq Y_n \leq \theta + \epsilon] = 1, \quad for\ every\ \epsilon > 0.$$

With this terminology, statement (26.7) is then the same thing as saying that the sample mean converges stochastically to μ (in the limit as the sample size becomes infinite).

Proof of (26.7). Let the positive number ϵ in statement (26.7) be chosen arbitrarily. Let C be any positive real number. If the sample size n is large enough, then

$$\epsilon \geq \frac{C\sigma}{\sqrt{n}},$$

from which it follows that the event

$$\{\mu - \epsilon \leq \bar{X}_n \leq \mu + \epsilon\} \tag{26.8}$$

contains the event

$$\left\{\mu - \frac{C\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + \frac{C\sigma}{\sqrt{n}}\right\}. \tag{26.9}$$

We learned in Chapter 1 that if an event E contains an event F, then $P(E) \geq P(F)$. Therefore, the probability of event (26.8) is $\geq$ the probability of event (26.9) if n is large enough. The CLT tells us that the probability of event (26.9) converges to

$$\int_{-C}^{C} f(z)dz,$$

where $f(z)$ is the standard Gaussian PDF. Therefore,

$$\lim_{n\to\infty} P[\mu - \epsilon \leq \bar{X}_n \leq \mu + \epsilon] \geq \int_{-C}^{C} f(z)dz \tag{26.10}$$

for every $C > 0$. As we make C extremely large, the right side of (26.10) becomes closer and closer to 1. Therefore, the limit on the side side of (26.10) has to be 1.

26.5 Confidence Interval Design: the Gaussian Case

Let the sampling distribution be Gaussian. Suppose the mean μ of our sampling distribution is unknown. Choose ϵ to be any positive real number that you want. (You can preset ϵ to be as close to zero as you want, such as $\epsilon = 0.1$ or $\epsilon = 0.01$.) Choose p to be any positive real number less than 1 that you want. (Typical choices of p might be $p = 0.90$ or $p = 0.95$.) Suppose you have found a sample siuze n for which

$$P\left[\bar{X}_n - \epsilon \leq \mu \leq \bar{X}_n + \epsilon\right] = p.$$

Then we call the interval of real numbers

$$[\bar{X}_n - \epsilon, \bar{X}_n + \epsilon] \tag{26.11}$$

a $100p\%$ *confidence interval for* μ. For example, if $p = 0.90$, we would call the interval (26.11) a 90% confidence interval for μ, meaning that if we re-compute the sample mean $\bar{X}_n$ on trial after trial for a large number of trials, approximately 90% of these trials will yield interval (26.11) which contains the unknown μ.

Useful Result: If we are sampling from a Gaussian distribution, then for every possible sample size n,

$$\left[\bar{X}_n - \frac{1.645\sigma}{\sqrt{n}},\ \bar{X}_n + \frac{1.645\sigma}{\sqrt{n}}\right]$$

is a 90% confidence interval for μ. We can state this result more compactly by saying that $\bar{X}_n \pm \frac{1.645\sigma}{\sqrt{n}}$ are the endpoints of a 90% confidence interval for μ.

Proof of Result. We must find the positive constant C such that

$$P\left[\bar{X}_n - \frac{C\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n - \frac{C\sigma}{\sqrt{n}}\right] = 0.90.$$

We can re-write the left side as

$$P\left[\mu - \frac{C\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu - \frac{C\sigma}{\sqrt{n}}\right]$$

We know from our earlier work with the Gaussian sampling distribution that this probability does not depend on μ, σ, or n. Therefore, this probability must be

$$P[-C \leq Z \leq C],$$

where Z is a standard Gaussian RV. (This is what you get when $n = 1$, $\mu = 0$, and $\sigma = 1$.) We have

$$P[-C \leq Z \leq C] = \Phi(C) - \Phi(-C) = 0.90,$$

where Φ is the standard Gaussian CDF. Use the fact that

$$\Phi(-C) = 1 - \Phi(C).$$

Simplifying, you conclude that

$$\Phi(C) = 0.95,$$

and then from table on page 123 of your textbook you conclude that $C = 1.645$.

Exercise. By a similar technique, prove that for all n,

$$\bar{X}_n \pm \frac{1.96\sigma}{\sqrt{n}}$$

are the endpoints of a 95% confidence interval for μ, when you are sampling from a Gaussian distribution.

26.6 Multivariate Density Example

Occasionally I will pause in our statistics coverage to go back to Chapter 5 and further our study of multivariate distributions. Let us consider the following example. Let X_1, X_2, X_3 be jointly continuously distributed RV's with multivariate density

$$f(x_1, x_2, x_3) = 1, \;\; (x_1, x_2, x_3) \in R \;\; (zero\ elsewhere),$$

where R is the unit cube

$$R = \{(x_1, x_2, x_3) : 0 \leq x_1 \leq 1;\; 0 \leq x_2 \leq 1;\; 0 \leq x_3 \leq 1\}.$$

Let's answer the following:

(a): Compute $P[X_1^2 + X_2^2 + X_3^2 \leq 1]$.

(b): Compute $P[X_1 + X_2 + X_3 \leq 1]$.

(c): Are X_1, X_2, X_3 independent?

Solution to (a). Let S be the three-dimensional region

$$S = \{(x_1, x_2, x_3) : x_1^2 + x_2^2 + x_3^2 \leq 1\} \cap R.$$

Then

$$P[X_1^2 + X_2^2 + X_3^2 \leq 1] = \iiint_S (1) dx_1 dx_2 dx_3 = volume(S).$$

It is not too hard to see that S is one-eighth of a sphere of radius one. The volume of a sphere of radius r is $\frac{4}{3}\pi r^3$. Plugging in $r = 1$ and taking $1/8$ of this, we conclude that

$$P[X_1^2 + X_2^2 + X_3^2 \leq 1] = \pi/6.$$

Solution to (b). Let S be the three-dimensional region

$$S = \{(x_1, x_2, x_3) : x_1 + x_2 + x_3 \leq 1\} \cap R.$$

Similarly to part(a), the probability we want is the volume of S. The required triple integral would extend over that part of the cube R lying below the plane $x_1 + x_2 + x_3 = 1$. One can see that this is the following, after maybe referring to your calculus book for some setups of limits on triple integrals over 3-D regions, in case you need to refresh your memory about how this is done:

$$volume(S) = \int_0^1 \int_0^{1-x_1} \int_0^{1-x_1-x_2} dx_3 dx_2 dx_1$$

The following Matlab script accomplishes this integration task:

```
syms x1 x2 x3
int(int(int(1,x3,0,1-x1-x2),x2,0,1-x1),0,1)

ans =

1/6
```

We conclude that

$$P[X_1 + X_2 + X_3 \leq 1] = 1/6.$$

Solution to (c). Independence means that

$$f(x_1, x_2, x_3) = f_{X_1}(x_1) f_{X_2}(x_2) f_{X_3}(x_3).$$

To find the marginal densities, you integrate out all the remaining variables from the multivariate density:

$$\begin{aligned} f_{X_1}(x_1) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3) dx_2 dx_3 \\ f_{X_2}(x_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3) dx_1 dx_3 \\ f_{X_3}(x_3) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3) dx_1 dx_2 \end{aligned}$$

You can just find one of these and then use symmetry to obtain the other two. You get

$$\begin{aligned} f_{X_1}(x_1) &= 1, \ \ 0 \leq x_1 \leq 1 \ (zero \ elsewhere) \\ f_{X_2}(x_2) &= 1, \ \ 0 \leq x_2 \leq 1 \ (zero \ elsewhere) \\ f_{X_3}(x_3) &= 1, \ \ 0 \leq x_3 \leq 1 \ (zero \ elsewhere) \end{aligned}$$

The product of these is clearly $f(x_1, x_2, x_3)$. The X_i's are indeed independent.

Remark. In terms of our statistics coverage, we can say that the RV's X_1, X_2, X_3 of this example form a random sample of size 3 from the Uniform$(0, 1)$ distribution.

Lecture 27

Statistics Part 2

27.1 Chebyshev's Inequality

You can find material on Chebyshev's Inequality on page 278 of your textbook. Chebyshev's inequality for any RV Y says that

$$P\left(\mu_Y - k\sigma_Y < Y < \mu_Y + k\sigma_Y\right) \geq 1 - \frac{1}{k^2}, \tag{27.1}$$

where k is any real number ≥ 1. I will give the simple proof of Chebyshev's Inequality at the end of this section.

To illustrate Chebyshev's Inequality, suppose we take $k = 2$. Then Chebyshev's inequality says that

$$P\left(\mu_Y - 2\sigma_Y < Y < \mu_Y + 2\sigma_Y\right) \geq 3/4 = 0.75.$$

Thus, no matter what the random variable, you are guaranteed to be within two standard deviations of the mean at least 75% of the time. Or, if $k = 3$, Chebyshev's inequality says

$$P\left(\mu_Y - 3\sigma_Y < Y < \mu_Y + 3\sigma_Y\right) \geq 8/9 = 0.889.$$

Thus, no matter what the random variable, you are guaranteed to be within three standard deviations of the mean at least 88% of the time.

Since the Chebyshev bound is valid for ALL RV's, the actual probability may be somewhat bigger than the Chebyshev bound for certain RV's (that is, the Chebyshev bound will not be very tight).

Example 27.1. Let us see what Chebyshev's Inequality says when the RV Y is Uniform(0,1) and $k = 1.5$. We have (see Appendix A if necessary):

$$\mu_Y \quad = \quad 1/2$$

$$\begin{aligned} \sigma_Y &= \sqrt{1/12} \\ \mu_Y - (1.5)\sigma_Y &= 0.0670 \\ \mu_Y + (1.5)\sigma_Y &= 0.9330 \end{aligned}$$

The exact probability of being within 1.5 standard deviations of the mean is

$$P[\mu_Y - (1.5)\sigma_Y < Y < \mu_Y + (1.5)\sigma_Y] = P[.0670 \leq Y \leq 0.9330] = 0.8660.$$

The Chebyshev lower bound is

$$1 - \frac{1}{k^2} = 1 - (1.5)^{-2} = 5/9 = 0.5556.$$

Notice that 0.8660 is considerably bigger than 0.5556. Therefore in this case the Chebyshev lower bound is not very tight.

Exercise. For Y Gaussian, compute the exact probability

$$P[\mu_Y - (1.5)\sigma_Y < Y < \mu_Y + (1.5)\sigma_Y]$$

using page 123 of your textbook and see how close this is to the Chebyshev lower bound 5/9.

Proof of Chebyshev's Inequality. Let us abbreviate μ_Y as μ and abbreviate σ_Y as σ. We start with the statement that

$$P[|Y - \mu| \geq k\sigma] = \int_{-\infty}^{\mu - k\sigma} f_Y(y)dy + \int_{\mu + k\sigma}^{\infty} f_Y(y)dy$$

In both integrals on the right, the inequality

$$k^2\sigma^2 \leq (y - \mu)^2$$

holds for all y in the range of integration. Therefore

$$\begin{aligned} \int_{-\infty}^{\mu - k\sigma} f_Y(y)dy &= k^{-2}\sigma^{-2} \int_{-\infty}^{\mu - k\sigma} k^2\sigma^2 f_Y(y)dy \leq k^{-2}\sigma^{-2} \int_{-\infty}^{\mu - k\sigma} (y - \mu)^2 f_Y(y)dy \\ \int_{\mu + k\sigma}^{\infty} f_Y(y)dy &= k^{-2}\sigma^{-2} \int_{\mu + k\sigma}^{\infty} k^2\sigma^2 f_Y(y)dy \leq k^{-2}\sigma^{-2} \int_{\mu + k\sigma}^{\infty} (y - \mu)^2 f_Y(y)dy \end{aligned}$$

Adding these two inequalities together, we obtain

$$\begin{aligned} P[|Y - \mu| \geq k\sigma] &\leq k^{-2}\sigma^{-2} \int_{-\infty}^{\mu - k\sigma} (y - \mu)^2 f_Y(y)dy + k^{-2}\sigma^{-2} \int_{\mu + k\sigma}^{\infty} (y - \mu)^2 f_Y(y)dy \\ &\leq k^{-2}\sigma^{-2} \int_{-\infty}^{\infty} (y - \mu)^2 f_Y(y)dy = k^{-2}\sigma^{-2}\sigma^2 = k^{-2} \end{aligned}$$

Taking the complement of the event $\{|Y - \mu| \geq k\sigma\}$, we obtain the event

$$\{\mu - k\sigma < Y < \mu + k\sigma\} \tag{27.2}$$

which must have probabilitity at least as big as $1 - (1/k)^2$. The event

$$\{\mu - k\sigma \leq Y \leq \mu + k\sigma\}$$

is even bigger than the event (27.2), and so it must have probability which is also at least as big as $1 - (1/k)^2$. In other words, we have proved (27.1).

27.2 Completion of Confidence Interval Design

In Section 26.5, we considered Confidence Interval Design for sampling from a Gaussian distribution. In this section, we complete our coverage of Confidence Interval Design by:

- explaining how to design confidence intervals when you sample from a nonGaussian distribution;
- showing how to determine how many samples n are needed for your sample of size n if you want a certain type of confidence interval, both in the case of a Gaussian smpling distribution and in the case of a nonGaussian sampling distribution.

27.2.1 NonGaussian Confidence Interval Design

As in Lecture 26, suppose you have $X_1, X_2, \cdots, X_n$, a sample of size n from your sampling distribution, where this distribution has unknown mean μ and known variance σ^2. If the sampling distribution is Gaussian, we showed in Section 26.5 how to find a constant k such that

$$\left[\bar{X}_n - \frac{k\sigma}{\sqrt{n}},\ \bar{X}_n + \frac{k\sigma}{\sqrt{n}}\right] \tag{27.3}$$

is a 100p% confidence interval for μ, for whatever p you want to set out in advance.

We now assume that the sampling distribution is nonGaussian. We again have to explain how to find k so that we obtain a confidence interval of the form (27.3) with a desired level of confidence. Suppose we let Y be equal to the sample mean $\bar{X}_n$ in Chebyshev's Inequality (27.1):

$$P\left[\mu_{\bar{X}_n} - k\sigma_{\bar{X}_n} \leq \bar{X}_n \leq \mu_{\bar{X}_n} + k\sigma_{\bar{X}_n}\right] \geq 1 - \frac{1}{k^2}.$$

Substituting

$$\begin{aligned} \mu_{\bar{X}_n} &= \mu \\ \sigma_{\bar{X}_n} &= \frac{\sigma}{\sqrt{n}}, \end{aligned}$$

we see that

$$P\left[\mu - \frac{k\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + \frac{k\sigma}{\sqrt{n}}\right] \geq 1 - \frac{1}{k^2}.$$

The event on the left side is unchanged if you exchange μ and $\bar{X}_n$, which gives us

$$P\left[\bar{X}_n - \frac{k\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{k\sigma}{\sqrt{n}}\right] \geq 1 - \frac{1}{k^2}.$$

We can now make the following immediate conclusion.

Conclusion: For any sampling distribution, the confidence interval (27.3) has percentage level of confidence at least $100(1 - \frac{1}{k^2})\%$.

Example 27.2. Suppose we want at least a 90% confidence intervals for μ. Then, by the Conclusion, we can choose k in the confidence interval (27.3) by solving the equation

$$100\left(1 - \frac{1}{k^2}\right) = 90,$$

which gives

$$k = \sqrt{10} = 3.1623,$$

to four decimal places. We conclude that no matter what distribution we sample from,

$$\left[\bar{X}_n - \frac{3.1623\sigma}{\sqrt{n}},\ \bar{X}_n + \frac{3.1623\sigma}{\sqrt{n}}\right] \tag{27.4}$$

is at least a 90% confidence interval for μ. Compare this with our earlier result for sampling from a Gaussian distribution. In this case, we determined that

$$\left[\bar{X}_n - \frac{1.645\sigma}{\sqrt{n}},\ \bar{X}_n + \frac{1.645\sigma}{\sqrt{n}}\right] \tag{27.5}$$

is a 90% confidence interval for μ. Notice that the nonGaussian confidence interval (27.4) is a bit wider than Gaussian confidence interval (27.5), which is less desirable since for the same level of confidence, a shorter confidence interval would be preferable to a longer one. This is the price we pay for our ignorance as to what type of distribution we are sampling from.

Exercise. For an arbitrary sampling distribution, prove that

$$\left[\bar{X}_n - \frac{4.4721\sigma}{\sqrt{n}},\ \bar{X}_n + \frac{4.4721\sigma}{\sqrt{n}}\right]$$

is at least a 95% confidence interval for μ.

27.2.2 How Many Samples?

In our preceding Confidence Interval Design discussions, we only concentrated on the percentage level of confidence of the confidence interval to be designed. If we also specify how wide the confidence interval should be, then we can pin down how many samples n we need to take in our random sample in order that the confidence interval we design will have the desired percentage level of confidence and the desired width. The following examples illustrate this.

Example 27.3. We sample from a Gaussian distribution with unknown mean μ and $\sigma = 1$. We want a 90% confidence interval

$$[\bar{X}_n - \epsilon,\ \bar{X}_n + \epsilon]$$

in which $\epsilon = 0.01$. Let us determine how big a sample size n we need in order to accomplish these design goals. First of all, we know that the 90% confidence interval for sampling from Gaussian distribution takes the form

$$\left[\bar{X}_n - \frac{1.645\sigma}{\sqrt{n}},\ \bar{X}_n + \frac{1.645\sigma}{\sqrt{n}}\right]$$

So, we set

$$\frac{1.645\sigma}{\sqrt{n}} = \epsilon = 0.01,$$

and solve for n with $\sigma = 1$. You will see that n will be about 27600. Thus, we need sample size about $n = 27600$ to achieve our goal in confidence interval design.

Example 27.4. Now, we sample from a nonGaussian distribution with unknown mean μ and $\sigma = 1$. We want a 90% confidence interval

$$[\bar{X}_n - \epsilon,\ \bar{X}_n + \epsilon]$$

in which $\epsilon = 0.01$. From Example 27.2, we know that $k = \sqrt{10}$ is the proper choice of k to determine the desired confidence level of at least 90%. We then find n by solving

$$\frac{k\sigma}{\sqrt{n}} = \frac{\sqrt{10}\sigma}{\sqrt{n}} = \epsilon = 0.01$$

with $\sigma = 1$. You see that $n = 100000$ is the sample size that is necessary. Note that this is a bit more than the 27600 samples needed for the Gaussian confidence interval. We have to take many more samples as a penalty for our ignorance concerning what the sampling distribution is.

Example 27.5. Suppose we now sample from a binary probability distribution with unknown mean μ. This binary distribution assigns probability p and $1 - p$ to the binary values 1 and 0, respectively, where p is an unknown paramter. The mean μ of this distribution is then seen to be p:

$$\mu = p * 1 + (1 - p) * 0 = p.$$

We leave it as an exercise for the reader to show that the standard deviation σ of this distribution is given by

$$\sigma = \sqrt{p(1-p)}.$$

Let us try to find how many samples n we need to obtain a 90% confidence interval

$$[\bar{X}_n - \epsilon, \bar{X}_n + \epsilon] \tag{27.6}$$

with $\epsilon = 0.01$. The tricky part of this problem is that σ depends on p and is therefore unknown. So the confidence interval cannot be of the form

$$\left[\bar{X}_n - \frac{k\sigma}{\sqrt{n}},\ \bar{X}_n + \frac{k\sigma}{\sqrt{n}}\right] \tag{27.7}$$

because the endpoints of this interval depend on the unknown σ. Here is how we can get around this difficulty. First, as in Examples 27.4. and 27.2, we can argue that if $k = \sqrt{10}$, then the interval (27.7) contains $\mu = p$ with probability at least 0.90. The resulting interval (27.7) varies with p and one really has infinitely many such intervals, one for each value of p. At this point, one can take the biggest of these intervals as the interval (27.6). This will tell us what n is and will also guarantee at least 90% confidence level (the biggest of the intervals has confidence level at least as great as any of the separate intervals—increasing the size of an interval makes the confidence level go up). Let us now go ahead and perform this procedure: plugging $k = \sqrt{10}$ in (27.7), you get the interval

$$\left[\bar{X}_n - \frac{\sqrt{10}\sqrt{p(1-p)}}{\sqrt{n}},\ \bar{X}_n + \frac{\sqrt{10}\sqrt{p(1-p)}}{\sqrt{n}}\right], \tag{27.8}$$

which is at least a 90% interval, although it depends on p. The biggest of these intervals, as p varies from $p = 0$ to $p = 1$, is for $p = 1/2$, because the max value of $\sqrt{p(1-p)}$ takes place at $p = 1/2$. We set $p = 1/2$ in (27.8) to see that this biggest interval is

$$\left[\bar{X}_n - \frac{\sqrt{10}\sqrt{1/4}}{\sqrt{n}},\ \bar{X}_n + \frac{\sqrt{10}\sqrt{1/4}}{\sqrt{n}}\right]$$

We want this interval to be the confidence interval (27.6), so we set

$$\epsilon = \frac{\sqrt{10}\sqrt{1/4}}{\sqrt{n}} = 0.01,$$

obtaining $n = 25000$ as the necessary number of samples. Our conclusion is that

$$[\bar{X}_n - 0.01,\ \bar{X} + 0.01]$$

is at least a 90% confidence interval for μ, if $n = 25000$, that is,

$$P[\bar{X}_n - 0.01 < \mu < \bar{X}_n + 0.01] \geq 0.90$$

for $n = 25000$, regardless of the value of $\mu = p$.

27.3 Multivariate Gaussian density example

Let RV's X_1, X_2, X_3 have the multivariate Gaussian density

$$f(x_1, x_2, x_3) = C \exp\left[-0.5(x_1\ x_2\ x_3)C_X^{-1}(x_1\ x_2\ x_3)^T\right], \tag{27.9}$$

where the covariance matrix C_X of the X_i's is

$$C_X = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

Let us work out the following:

(a): Find $f_{X_1,X_3}(x_1, x_3)$.

(b): Find $f_{X_2}(x_2)$

Solution to (a). The hard way to find the answer would be to do the integral

$$f_{X_1,X_3}(x_1, x_3) = \int_{-\infty}^{\infty} f(x_1, x_2, x_3)dx_2.$$

Fortunately, we can use a property of multivariate Gaussian distributions to obtain the answer in a much simpler manner. This property states that any subset of multivariate Gaussian RV's is also multivariate Gaussian. Therefore, X_1, X_3 automatically have a bivariate Gaussian density. To express this density, we need the means of X_1, X_3 and the covariance matrix for X_1, X_3. From the form of $f(x_1, x_2, x_3)$ in (27.9), it is clear that the means of X_1, X_2, X_3 are all zero. (If one of these means were nonzero, there would be at least one linear term in the x_i's in the exponent on the right side of (27.9), but there are only quadratic terms.) Therefore, X_1 and X_3 both have mean 0. The covariance matrix for X_1, X_3 is the 2×2 matrix consisting of the four corner elements of C_X, namely, the matrix

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

The inverse of this matrix is

$$\begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix}.$$

We now form the triple product

$$(x_1\ x_3)\begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix}(x_1\ x_3)^T,$$

which multiplies out as

$$(2/3)x_1^2 + (2/3)x_3^2 - (2/3)x_1x_3.$$

If we multiply this by $-1/2$, we obtain the exponent of the $f_{X_1,X_3}(x_1, x_3)$ density:

$$\begin{aligned} f_{X_1,X_3}(x_1, x_3) &= C' \exp[-0.5\{(2/3)x_1^2 + (2/3)x_3^2 - (2/3)x_1x_3\}]. \\ &= C' \exp[-(x_1^2 + x_3^2 - x_1x_3)/3] \end{aligned}$$

From page 191 of your textbook, the constant C' is given by

$$C' = \frac{1}{2\pi\sigma_{X_1}\sigma_{X_3}\sqrt{1-\rho_{X_1,X_3}^2}} = \frac{1}{2\pi\sqrt{3}}.$$

(Obviously, $\sigma_{X_1} = \sigma_{X_3} = 2$. We leave it to the reader to show that $\rho_{X_1,X_3} = 1/2$.)

Solution to (b). X_2 must be Gaussian. Its mean is zero. Its variance is 2 (the middle element of C_X). Therefore, its density (from App A) is

$$f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}\sqrt{2}} \exp\left(-\frac{x_2^2}{4}\right).$$

This is much easier than doing the integral

$$f_{X_3}(x_3) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2, x_3)dx_1dx_3.$$

Lecture 28

Statistics Part 3

28.1 Confidence Interval for $P(E)$

Suppose we have an event E associated with a random experiment whose probability $P(E)$ is unknown. Going back to the beginnings of EE 3025, we estimated $P(E)$ empirically as follows:

- Perform the experiment n times (independent trials).
- Of these n trials, let n_E be the total number of these trials on which E occurred.
- Use the "relative frequency" n_E/n as an estimate for $P(E)$.

It stands to reason that we may be able to form a confidence interval for $P(E)$ of the form

$$\left[\frac{n_E}{n} - \epsilon,\ \frac{n_E}{n} - \epsilon\right]. \tag{28.1}$$

We will want a certain level of confidence for this interval, namely, for some preselected α strictly between 0 and 1, we will want to ensure that

$$P\left[\frac{n_E}{n} - \epsilon \leq P(E) \leq \frac{n_E}{n} + \epsilon\right] \geq \alpha. \tag{28.2}$$

We may state our confidence interval design problem as follows:

Conf Int Design Problem: Given α and ϵ, find a number of samples that will ensure that inequality (28.2) is true.

First, I state how you find n in order to solve this problem. Then, I explain why my solution is valid.

Solution to Conf Int Design Problem: First, find k so that

$$1 - \frac{1}{k^2} = \alpha. \tag{28.3}$$

Then, solve the following equation for n:

$$\frac{k}{2\sqrt{n}} = \epsilon. \tag{28.4}$$

If n is a positive integer, this is the number of trials you take to form the confidence interval (28.2). If not, you round up to the smallest integer greater than or equal to the solution to (28.4).

Example 28.1. Let us figure out how many trials n will be enough so that (28.1) will be a 90% confidence confidence interval with $\epsilon = 0.01$. Then, $\alpha = 0.90$. Solving (28.3) for k, we obtain $k = \sqrt{10}$. Solving (28.4) with $\epsilon = 0.01$, we see that the number of trials n to be used in our confidence interval is

$$n = 25000.$$

The design of our confidence interval is now complete. In other words, if you repeatedly take 25000 trials and compute the endpoints of the confidence interval (28.1), you will find that in the long run at least 90% of these confidence intervals will contain $P(E)$, no matter what the value of $P(E)$ actually is.

Example 28.2. Figure out how many trials n will be enough so that (28.1) will be a 95% confidence confidence interval with $\epsilon = 0.01$.

Proof that our design method works. Suppose you take a random sample of size n

$$X_1, X_2, \cdots, X_n$$

from the binary probability distribution which assigns probability $p = P(E)$ to the binary value 1 and then it is automatically true that probability $1 - p$ is assigned to the binary value 0. Then, we can regard the sample mean $\bar{X}_n$ of this random sample as coinciding with the relative frequency n_E/n:

$$\frac{n_E}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \bar{X}_n. \tag{28.5}$$

(You can regard each X_i as taking the value 1 precisely when event E occurs, so that the sum of the X_i's is therefore n_E, the total number of times E occurs.) The mean μ of our binary prob dist is $\mu = p = P(E)$. Using the method of Example 27.5, one finds an interval of the form

$$\left[\bar{X}_n - \frac{k\sigma}{\sqrt{n}},\ \bar{X}_n + \frac{k\sigma}{\sqrt{n}}\right]$$

so that

$$P\left[\bar{X}_n - \frac{k\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{k\sigma}{\sqrt{n}}\right] \geq \alpha, \tag{28.6}$$

where σ is the standard deviation of our binary prob dist, given by

$$\sigma = \sqrt{p(1-p)}.$$

Using Chebyshev's Inequality, the k needed for (28.6) to be true is found by solving equation (28.3). Plugging this k value into (28.6) and also plugging in the σ value, we've shown (using (28.5)) that the following is true:

$$P = \left[\frac{n_E}{n} - \frac{k\sqrt{p(1-p)}}{\sqrt{n}} \leq P(E) \leq \frac{n_E}{n} + \frac{k\sqrt{p(1-p)}}{\sqrt{n}}\right] \geq \alpha$$

We need to find two endpoints not depending on $p = P(E)$. If we replace $p(1-p)$ in the two endpoints by the biggest it can be (namely, 1/4), we get a bigger interval for which the preceding prob statement is still true. Our conclusion is that the following is true for k satisfying (28.3):

$$P\left[\frac{n_E}{n} - \frac{k}{2\sqrt{n}} \leq P(E) \leq \frac{n_E}{n} + \frac{k}{2\sqrt{n}}\right] \geq \alpha$$

Solving the equation (28.4), we obtain the necessary value of n for our confidence interval (28.1).

28.2 Application to Hypothesis Testing

(Note: You are not responsible for this section on EE 3025 Exam 2.)

Chapter 8 of your textbook is on hypothesis testing. Although hypothesis testing is very important, there was no time to cover Chapter 8 in EE 3025. However, it is interesting to note that the simplest hypothesis testing problem is solvable using the confidence interval approach. This section is devoted to showing you this.

Suppose you have a probability distribution with unknown mean μ and known variance σ^2. To make the problem of estimating μ simpler, suppose you do know that there are exactly two possibilities for μ, namely, μ can either be equal to a known value μ_0 or else μ can be equal to a known value μ_1 (let us assume that $\mu_1 > \mu_0$). Statisticians would state these two possibilities as two "hypotheses": Either the hypothesis

$$H_0 : \mu = \mu_0$$

is true, or else the hypothesis

$$H_1 : \mu = \mu_1$$

is true.

Hypothesis Testing Problem: Based an a random sample of size n from the underlying prob dist, decide whether hypothesis H_0 is true or whether hypothesis H_1 is true.

In order to solve the Hypothesis Testing Problem, we need to formulate a decision rule that will tell us exactly when to decide that H_0 is true and when to decide that H_1 is true. Here is one possible decision rule, based on confidence intervals:

Decision Rule: Suppose you want to make the correct decision at least 90% of the time about which of the two hypotheses H_0, H_1 is true, no matter which one is actually true. Set

$$\epsilon = \frac{|\mu_1 - \mu_0|}{2}, \tag{28.7}$$

and find the sample size n needed so that

$$[\bar{X}_n - \epsilon,\ \bar{X}_n + \epsilon] \tag{28.8}$$

will be a 90% confidence interval for μ. Take your random sample of size n and compute the endpoints of your confidence interval. Make your decision about the hypotheses as follows:

- If your confidence interval (28.8) contains μ_1, then decide that hypothesis H_1 is true. (This decision makes sense because if your interval contains μ_1, it cannot also contain μ_0, because of the choice of ϵ in (28.7).)
- If your confidence interval (28.8) contains μ_0, then decide that hypothesis H_0 is true. (This decision makes sense because if your interval contains μ_0, it cannot also contain μ_1.)
- If your confidence interval contains neither μ_0 nor μ_1, then you can just make a guess as to which of the two hypotheses is true.

Discussion. We discuss why the preceding Decision Rule will make the correct decision at least 90% of the time (in the long run as you re-compute your confidence interval over and over again and make the resulting repeated decisions). By the fact that the confidence interval is a 90% confidence interval, it will contain the actual value of μ at least 90% of the time in the long run. But μ can only be μ_0 or μ_1 and nothing else. If $\mu = \mu_0$ is true, then since the confidence interval contains μ at least 90% of the time, it will contain μ_0 at least 90% of the time, and on such occasions, the decision rule will always select hypothesis H_0 as the true hypothesis (which is the correct decision). On the other hand, if $\mu = \mu_1$ is true, then since the confidence interval contains μ at least 90% of the time, it will contain μ_1 at least 90% of the time, and on such occasions, the decision rule will always select hypothesis H_1 as the true hypothesis (which is the correct decision). We conclude that the correct decision will be made at least 90% of the time as to which hypothesis is true, no matter which of the two hypotheses is actually true. (For those occasions in which the confidence

interval contains neither μ_0 nor μ_1, the decision rule might guess that the wrong hypothesis is true, but this eventuality can only occur less than 10% of the time.)

Remark. If you want your decision rule for the hypothesis testing problem to be correct more than 90% of the time, then you could design your confidence interval to be a 95% confidence interval or a 97.5% confidence interval, or however close to 100% confidence level that you want. The percentage level of confidence of your confidence interval determines the percentage of time that your decision rule will be correct.

28.3 Linear Transformation of Multivariate Densities

(Note: You are not responsible for this section on EE 3025 Exam 2.)

Suppose you have jointly continuous RV's $X_1, X_2, \cdots, X_n$. Then they have a multivariate density

$$f(x_1, x_2, \cdots, x_n)$$

Suppose you linearly transform these "old RV's" to get "new RV's" $Y_1, Y_2, \cdots, Y_n$. We can express the linear transformation in matrix format as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = A \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix},$$

where A is an $n \times n$ invertible matrix with constant entries. The new RV's have a joint density

$$g(y_1, y_2, \cdots, y_n).$$

The question we want to consider in this section is how to compute the density $g(y_1, y_2, \cdots, y_n)$ from the density $f(x_1, x_2, \cdots, x_n)$. Here is a formula for doing this, from Theorem 5.11 (page 223) of your textbook:

$$g(y_1, y_2, \cdots, y_n) = \left(\frac{1}{|\det(A)|} \right) f(x_1, x_2, \cdots, x_n), \tag{28.9}$$

where in the right hand side we substitute for each x_j variable in terms of the y_i's. I will not prove formula (28.9) here, because it is a special case of a formula I will give in a future lecture for transforming a multivariate density under a *nonlinear* transform. Intuitively, however, one can "psych out" why the new multivariate density would just be a constant multiple of the old multivariate density (at least in two dimensions): A linear transformation in 2-D transforms parallelograms in the x_1, x_2 plane into parallelograms in the y_1, y_2 plane, where the area of the new parallelogram is always just a fixed constant multiple of the area of the old parallelogram. In proving formula (28.9) in 2-D, you would compute multivariate probability over a new region in terms of multivariate

probability over an old region by approximating the new and old regions as unions of infinitesimally small new and old parallograms, respectively. The constant scaling factor between the areas of the corresponding new and old infinitesimal parallelograms would yield, in the limit, expressions for the new and old multivariate probs as 2-D integrals with 2-D density function integrands that are also related by this constant multiple.

Example 28.2. Let X_1, X_2 be independent RV's. Make the change of variable

$$\begin{aligned} Y_1 &= X_1 + X_2 \\ Y_2 &= X_2 \end{aligned}$$

Let's use formula (28.9) to express $f_{Y_1,Y_2}(y_1, y_2)$ in terms of $f_{X_1,X_2}(x_1, x_2)$. The coefficient matrix of the transformation is

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

which has determinant 1. Also, if you solve for the old variables in terms of the new variables, you get

$$\begin{aligned} X_1 &= Y_1 - Y_2 \\ X_2 &= Y_2 \end{aligned}$$

We conclude from (28.9) that

$$f_{Y_1,Y_2}(y_1, y_2) = (1) f_{X_1,X_2}(x_1, x_2) = f_{X_1,X_2}(y_1 - y_2, y_2).$$

Up to now, I have not used the fact that RV's X_1, X_2 are independent. If I now use that fact, I obtain

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1}(y_1 - y_2) f_{X_2}(y_2).$$

As a byproduct, let us now find the PDF of Y_1. We obtain:

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f_{X_1}(y_1 - y_2) f_{X_2}(y_2) dy_2.$$

The right side of the preceding equation is your old friend the *convolution integral* from EE 3015. We conclude that

$$f_{Y_1} = f_{X_1} * f_{X_2}.$$

In other words, if you sum up two independent RV's X_1, X_2, the PDF of the sum RV is the convolution of the separate PDF's of X_1 and X_2. This is a result that we have been using for some time. We now see why it is true.

Example 28.3. Suppose X_1, X_2 have joint density

$$f(x_1, x_2) = \left(\frac{1}{2\pi\sqrt{3}}\right) \exp[-0.5\{(2/3)x_1^2 + (2/3)x_2^2 - (2/3)x_1 x_2\}]. \tag{28.10}$$

(This is the bivariate Gaussian density that appeared toward the end of Section 27.3.) The purpose of this example is to point out that I can linearly transform the RV's X_1, X_2 to obtain independent Gaussian(0,1) RV's Y_1, Y_2. To do this, I first rewrite the exponent in the right side of (28.10) as

$$\exp[-0.5\ (x_1\ x_2) \begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix} (x_1\ x_2)^T].$$

To see what linear transformation is needed, we first need to find the covariance matrix C_X of the X_i's. This is obtained by inverting the 2×2 matrix that appears in the middle of the exponent:

$$\begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix}^{-1} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = C_X.$$

Now choose a 2×2 invertible matrix A so that

$$AC_XA^T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{28.11}$$

the 2×2 identity matrix. There are infinitely many different A matrices that will satisfy the equation (28.11). In a good linear algebra course, you would learn at least one or two good techniques for finding a solution. To finish up, all I need is just one solution. You can easily verify that the following is a solution:

$$A = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{6} & 1/\sqrt{6} \end{pmatrix}.$$

The change of variable we will make is then

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = A \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}. \tag{28.12}$$

The reciprocal of the determinant of A is $\sqrt{3}$. Therefore, by formula (28.9), the new joint density is

$$f_{Y_1,Y_2}(y_1, y_2) = \sqrt{3} \left(\frac{1}{2\pi\sqrt{3}}\right) \exp[-0.5\ (x_1\ x_2)C_X^{-1}(x_1\ x_2)^T], \tag{28.13}$$

except that we still have to substitute for the old variables x_1, x_2 on the right side in terms of the new variables y_1, y_2. Solving equation (28.11) for C_X, we obtain

$$C_X = A^{-1}(A^T)^{-1},$$

and then inverting both sides of this equation we obtain

$$C_X^{-1} = A^TA.$$

The exponential part on the right side of (28.13) is then

$$\exp[-0.5\ (x_1\ x_2)A^T A(x_1\ x_2)^T],$$

which can be rewritten as

$$\exp[-0.5\ \{(x_1\ x_2)A^T\}\{(x_1\ x_2)A^T\}^T],$$

and then as

$$\exp[-0.5\ (y_1\ y_2)(y_1\ y_2)^T],$$

because from (28.12) we have

$$(y_1\ y_2) = (x_1\ x_2)A^T.$$

We conclude that our new joint PDF in simplest form is

$$f_{Y_1,Y_2}(y_1, y_2) = \left(\frac{1}{2\pi}\right) \exp[-0.5(y_1^2 + y_2^2)],$$

which factors as the product of two Gaussian(0,1) PDF's. We have shown that the change of variable (28.12) converts the old RV's X_1, X_2 into new RV's Y_1, Y_2 which are independent Gaussian(0,1) RV's.

Final Remark. The technique illustrated in the preceding example can be applied to any set $X_1, X_2, \cdots, X_n$ of multivariate Gaussian RV's having zero mean. You can linearly transform them into independent Gaussian(0,1) RV's $Y_1, Y_2, \cdots, Y_n$ if you use the appropriate $n \times n$ matrix A to convert the column vector $[X_1, X_2, \cdots, X_n]^T$ into the column vector $[Y_1, Y_2, \cdots, Y_n]^T$. Any $n \times n$ matrix A such that

$$AC_X A^T$$

is the $n \times n$ identity matrix will do the trick!

Lecture 29

Statistics Part 4

29.1 Straight Line Regression of the Mean

We start with an instructive example.

Example 29.1. Given X, Y satisfying:

- $\mu_X = 0$, $\mu_Y = 1$
- $Var(X) = 2$, $Var(Y) = 4$
- $\rho_{X,Y} = 1/2$

Assume that there are constants a, b such that

$$E[Y|X = x] = ax + b, \quad for\ all\ x.$$

This is called *straight line regression of the mean.* Use Law of Iterated Expectation to find a, b.

Solution.

$$E[Y] = E[E[Y|X]] = E[aX + b] = aE[X] + b = b.$$

Therefore $b = 1$.

$$E[XY] = E[XE[Y|X]] = E[X(aX + 1)] = aE[X^2] + E[X] = 2a.$$

Also,

$$E[XY] = Cov(X, Y) + \mu_X\mu_Y = Cov(X, Y) = \rho\sigma_X\sigma_Y = \sqrt{2}.$$

Therefore,

$$a = \sqrt{2}/2.$$

Using the Law of Iterated Expectation much as in Example 29.1, one can prove the following.

- If $E[Y|X=x]$ is a straight line function (i.e., of the form $ax+b$ for constants a,b), then it takes the unique form
$$E[Y|X=x] = \mu_Y + \frac{\rho_{X,Y}\sigma_Y}{\sigma_X}(x-\mu_X).$$
- If $E[X|Y=y]$ is a straight line function (i.e., of the form $cy+d$ for constants c,d), then it takes the unique form
$$E[X|Y=y] = \mu_X + \frac{\rho_{X,Y}\sigma_X}{\sigma_Y}(y-\mu_Y).$$

29.2 Bivariate Gaussian Distribution

We say that random variables X, Y have a joint Gaussian distribution if their joint PDF takes the form

$$\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2+\left(\frac{y-\mu_y}{\sigma_y}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right\}\right] \quad (29.1)$$

where $\mu_x, \mu_y, \sigma_x, \sigma_y, \rho$ can be any five parameter values which satisfy

$$\begin{aligned} -\infty < &\ \mu_x &< \infty \\ -\infty < &\ \mu_y &< \infty \\ 0 < &\ \sigma_x &< \infty \\ 0 < &\ \sigma_y &< \infty \\ -1 < &\ \rho &< 1 \end{aligned}$$

Properties of Bivariate Gaussian

We will prove the following properties of the Bivariate Gaussian in Section 29.2.3.

- The marginal distributions are Gaussian:
$$\begin{aligned} X &\sim Gaussian(\mu_x, \sigma_x^2) \\ Y &\sim Gaussian(\mu_y, \sigma_y^2) \end{aligned}$$
- The conditional distributions are Gaussian: The cond dist of X given $Y=y$ is Gaussian with
$$\begin{aligned} E[X|Y=y] &= \mu_x + \rho(\sigma_x/\sigma_y)(y-\mu_y) \quad (29.2) \\ Var[X|Y=y] &= \sigma_x^2(1-\rho^2) \end{aligned}$$

The cond dist of Y given $X = x$ is Gaussian with

$$\begin{aligned} E[Y|X = x] &= \mu_y + \rho(\sigma_y/\sigma_x)(x - \mu_x) \\ Var[X|Y = y] &= \sigma_y^2(1 - \rho^2) \end{aligned}$$

- $\rho_{x,y} = \rho$.

We have seen in the recitations that a Gaussian joint density $f(x, y)$ can be plotted as a surface $z = f(x, y)$ in an (x, y, z)-coordinate system. Geometrical properties of the surface $z = f(x, y)$ are related to properties of the joint Gaussian probability distribution. For example, the point in the (x, y) plane at which the surface $z = f(x, y)$ reaches its maximum is the point (μ_x, μ_y). Planes of the form $x = C$ cut the surface $z = f(x, y)$ in one-dimensional curves which, when scaled properly, yield the conditional densities of Y given values of X. Planes of the form $y = C$ cut the surface $z = f(x, y)$ in one-dimensional curves which yield the conditional densities of X given values of Y when scaled properly. Cross-sections of the surface $z = f(x, y)$ with planes of the form $z = C$ are ellipses.

29.2.1 Effect of the parameter ρ

- The jointly Gaussian random variables X, Y are independent if and only if $\rho = 0$. The reader can see that the joint Gaussian density in (29.1) factors in this case as

$$\left[\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right)\right]\left[\frac{1}{\sqrt{2\pi}\sigma_y}\exp\left(-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right)\right]$$

  If $\rho = 0$, and the variances of X and Y are equal, then the surface $z = f(x, y)$ has a symmetrical bell shape, with the cross sections of the surface parallel to the (x, y)-plane being circles.

- As $\rho \to \pm 1$, the elliptical cross-sections of the surface $z = f(x, y)$ become more and more eccentric, with the surface $z = f(x, y)$ looking more and more "squashed" along the major axis of these elliptical cross sections. In the limit when ρ becomes 1, one has

$$P[Y = \mu_y + (\sigma_y/\sigma_x)(X - \mu_x)] = 1,$$

  which means that (X, Y) is concentrated along a line of positive slope. In the limit when ρ becomes -1, then

$$P[Y = \mu_y - (\sigma_y/\sigma_x)(X - \mu_x)] = 1,$$

  meaning that (X, Y) is concentrated along a line of negative slope.

29.2.2 Effect of Linear Change of Variable

Let X, Y be jointly Gaussian. Suppose we define two new random variables U, V by

$$\begin{aligned} U &= AX + BY + C \\ V &= DX + EY + F \end{aligned}$$

where A, B, C, D, E, F are constants. To avoid a degenerate situation, we require that $AE - BD \neq 0$. Under such a linear change of variable, it turns out that the new random variables U, V are also jointly Gaussian. To find the joint density of (U, V), you'd just have to perform the following three steps

Step 1: Compute the values of the parameters

$$\mu_u, \mu_v, \sigma_u^2, \sigma_v^2, \sigma_{u,v}$$

from the values of the parameters

$$\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_{x,y}$$

using the equations

$$\begin{aligned} \begin{bmatrix} \mu_u \\ \mu_v \end{bmatrix} &= \begin{bmatrix} A & B \\ D & E \end{bmatrix} \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} + \begin{bmatrix} C \\ F \end{bmatrix} \\ \begin{bmatrix} \sigma_u^2 & \sigma_{u,v} \\ \sigma_{u,v} & \sigma_v^2 \end{bmatrix} &= \begin{bmatrix} A & B \\ D & E \end{bmatrix} \begin{bmatrix} \sigma_x^2 & \sigma_{x,y} \\ \sigma_{x,y} & \sigma_y^2 \end{bmatrix} \begin{bmatrix} A & B \\ D & E \end{bmatrix}^T \end{aligned}$$

Step 2: Compute $\rho_{u,v} = \sigma_{u,v}/(\sigma_u \sigma_v)$.

Step 3: Plug the parameter values $\mu_u, \mu_v, \sigma_u^2, \sigma_v^2, \rho_{u,v}$ into the following expression for the joint density of U and V:

$$\frac{1}{2\pi\sigma_u\sigma_v\sqrt{1-\rho_{u,v}^2}} \exp\left[-\frac{1}{2(1-\rho_{u,v}^2)}\left\{\left(\frac{u-\mu_u}{\sigma_u}\right)^2 + \left(\frac{v-\mu_v}{\sigma_v}\right)^2 - 2\rho_{u,v}\left(\frac{u-\mu_u}{\sigma_u}\right)\left(\frac{v-\mu_v}{\sigma_v}\right)\right\}\right]$$

29.2.3 Proofs

The "Properties of Bivariate Gaussian" given earlier can be established by proving them for the special case in which

$$\mu_X = \mu_Y = 0, \quad \sigma_X = \sigma_Y = 1. \tag{29.3}$$

Discussion. You obtain RV's as in (29.3) by the change of variable

$$\begin{aligned} X' &= \frac{X-\mu_X}{\sigma_X} \\ Y' &= \frac{Y-\mu_Y}{\sigma_Y} \end{aligned}$$

For example, if you've shown that X' is Gaussian(0,1), it will follow automatically that X is Gaussian(μ_X, σ_X^2). Similarly, other properties of (X, Y) follow from properties of (X', Y').

According to (29.3), we now consider the special case of a bivariate Gaussian distribution in which the random variables X, Y have the joint density

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\{x^2+y^2-2\rho xy\}\right) \tag{29.4}$$

The parameter ρ is selected so as to satisfy $-1 < \rho < 1$. We want to compute the following entities:

$$f_X(x), f_Y(y), f(x|y), f(y|x), r_{X,Y}, \sigma_{X,Y}, \rho_{X,Y}$$

We could compute all of these things by brute force integration. Instead, we will try to be a little more subtle, in order to simplify our work. First, by the method of "completing the square", one derives the identity

$$x^2+y^2-2\rho xy = (x-\rho y)^2 + (1-\rho^2)y^2$$

This allows us to break apart the exponent in (29.4) and then to factor (29.4) as

$$f(x,y) = \left[\frac{1}{\sqrt{2\pi}} \exp\left(-y^2/2\right)\right] \left[\frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\{x-\rho y\}^2\right)\right] \tag{29.5}$$

The first factor within brackets on the right side of (29.5) is a $N(0,1)$ density. The second factor, considered as a function of x for fixed y, is a Gaussian density with mean ρy and variance $1-\rho^2$. Integrating to get $f_Y(y)$, you get

$$f_Y(y) = [first\ factor] \int_{-\infty}^{\infty} [second\ factor]\, dx$$

Since the integral of the second factor is one (it is a density!), we see that $f_Y(y)$ is equal to the first factor, which then identifies the second factor for us as the conditional density $f(x|y)$. We conclude:

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}} \exp(-y^2/2) \\ f(x|y) &= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}\{x-\rho y\}^2) \end{aligned}$$

By symmetry (reversing the roles of x and y), we then conclude that

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \\ f(y|x) &= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}\{y-\rho x\}^2) \end{aligned}$$

Putting all of this together, we can say the following.

- X and Y are each standard Gaussian
- The conditional distribution of X given $Y = y$ is Gaussian with the conditional mean and variance

$$\begin{aligned} E[X|Y=y] &= \rho y \\ \mathrm{V}ar[X|Y=y] &= 1-\rho^2 \end{aligned}$$

- The conditional distribution of Y given $X = x$ is Gaussian with the conditional mean and variance

$$\begin{aligned} E[Y|X=x] &= \rho x \\ \mathrm{V}ar[Y|X=x] &= 1-\rho^2 \end{aligned}$$

We now determine $r_{X,Y}, \sigma_{X,Y}, \rho_{X,Y}$. Since the means of X and Y are zero, $r_{X,Y} = \sigma_{X,Y}$. Since the variances of X and Y are one, $\sigma_{X,Y} = \rho_{X,Y}$. Therefore, all three of these quantities are equal. Observe that

$$E[XY|Y=y] = E[Xy|Y=y] = yE[X|Y=y] = \rho y^2$$

Therefore,

$$\begin{aligned} r_{X,Y} = E[XY] &= \int_{-\infty}^{\infty} E[XY|Y=y] f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \rho y^2 f_Y(y) dy \\ &= \rho E[Y^2] = \rho \end{aligned}$$

We conclude that

$$r_{X,Y} = \sigma_{X,Y} = \rho_{X,Y} = \rho$$

29.3 Proof of CLT

We prove the Central Limit Theorem in a special case. The random sample $X_1, X_2, \cdots, X_n$ is drawn according to the density

$$(1/2)\delta(x+1) + (1/2)\delta(x-1)$$

The mean μ is 0 and the variance σ is 1. Thus, we may express our normalized sum Z_n as

$$Z_n = \frac{(X_1 + X_2 + \ldots X_n) - n\mu}{\sqrt{n}\sigma} = \frac{X_1 + X_2 + \ldots X_n}{\sqrt{n}},$$

and so we are to show that

$$\frac{X_1 + X_2 + \ldots X_n}{\sqrt{n}} \approx Gaussian(0,1),$$

for large n. Notice that

$$Z_n = \left(\frac{X_1}{\sqrt{n}}\right) + \left(\frac{X_2}{\sqrt{n}}\right) + \ldots + \left(\frac{X_n}{\sqrt{n}}\right) \tag{29.6}$$

From (29.6), we can deduce the following relationship among the moment generating functions involved:

$$M_{Z_n}(s) = [M_{X_1/\sqrt{n}}(s)]^n. \tag{29.7}$$

We have

$$M_{X_1/\sqrt{n}}(s) = (1/2)e^{s/\sqrt{n}} + (1/2)e^{-s/\sqrt{n}} = \cosh(s/\sqrt{n}),$$

and therefore (29.7) becomes

$$M_{Z_n}(s) = [\cosh(s/\sqrt{n}]^n.$$

To show that Z_n becomes closer and closer to having a Gaussian(0,1) distribution as n becomes large is equivalent to showing that the moment generating function $M_{Z_n}(s)$ becomes closer and closer to the moment generating function $\exp(s^2/2)$ of a Gaussian(0,1) RV. Thus, the CLT for our special case will be proved if we can show that

$$\lim_{n\to\infty} [\cosh(s/\sqrt{n}]^n = \exp(s^2/2).$$

Taking the natural log of both sides, this reduces to showing that

$$\lim_{n\to\infty} n \log_e(\cosh(s/\sqrt{n})) = s^2/2,$$

which, making the change of variable $\alpha = 1/\sqrt{n}$, is the same as showing that

$$\lim_{\alpha\to 0} \frac{\log_e(\cosh(\alpha s))}{\alpha^2} = s^2/2.$$

This last relationship can be shown by using L'Hospital's Rule twice: Using L'Hospital the first time, you get

$$\lim_{\alpha \to 0} \frac{\log_e(\cosh(\alpha s))}{\alpha^2} = s \lim_{\alpha \to 0} \frac{\sinh(\alpha s))}{2\alpha}.$$

Using L'Hospital the second time, you get

$$s \lim_{\alpha \to 0} \frac{\sinh(\alpha s))}{2\alpha} = s^2 \lim_{\alpha \to 0} \frac{\cosh(\alpha s))}{2} = s^2/2.$$

Lecture 30

Statistics Part 5

30.1 Review Example

Let X, Y have bivariate Gaussian joint density of form

$$f_{X,Y}(x,y) = C \exp\left[-0.5\{4x^2 - 16xy - 8x + 20y^2 - 16y\}\right]. \tag{30.1}$$

Find the cond mean $E[X|Y=y]$ and the cond var $Var[X|Y=y]$. I give two solutions.

Method 1: If you fix y in the joint density and then change the C in front to make the integral with respect to x equal to 1, then you obtain the cond PDF $f_{X|Y}(x|y)$. It will be a Gaussian cond PDF and therefore of the form

$$C' \exp\left[-0.5\left\{\frac{(x - condmean)^2}{condvar}\right\}\right].$$

The *condmean* term is of the form $ay + b$ and the *condvar* term is a constant; these will give us $E[X|Y=y]$ and $Var[X|Y=y]$, respectively. To get this form, we just have to complete the square on the x terms within braces in (30.1):

$$4x^2 - 16xy - 8x = 4[x^2 - (4y+2)x] = 4[(x - \{2y+1\})^2] + \phi(y),$$

where $\phi(y)$ is an additional term which we can ignore because it depends on y alone and we are holding y fixed. We immediately conclude that

$$\begin{aligned} E[X|Y=y] &= 2y+1 \\ Var[X|Y=y] &= 1/4 \end{aligned}$$

Method 2: We have

$$E[X|Y=y] = \mu_X + \frac{\rho\sigma_X}{\sigma_Y}(y-\mu_Y) \tag{30.2}$$

$$Var[X|Y=y] = \sigma_X^2(1-\rho^2) \tag{30.3}$$

If we can compute the 5 parameters μ_X, μ_Y, σ_X^2, σ_Y^2, and ρ, then we just plug these in the two preceding equations. The point $(x,y) = (\mu_X, \mu_Y)$ is where the Gaussian density surface $z = f_{X,Y}(x,y)$ reaches its peak value. This is therefore the point at which the quantity in braces in (30.1) is minimized, and we can find this point by solving the equations

$$\begin{aligned} \frac{\partial}{\partial x}\{4x^2 - 16xy - 8x + 20y^2 - 16y\} &= 0 \\ \frac{\partial}{\partial y}\{4x^2 - 16xy - 8x + 20y^2 - 16y\} &= 0 \end{aligned}$$

These equations simplify to

$$\begin{aligned} 8x - 16y &= 8 \\ -16x + 40y &= 16 \end{aligned}$$

Solving, we get

$$\mu_X = 9, \quad \mu_Y = 4$$

To find the three remaining parameters, just look at the quadratic terms within the braces in (30.1):

$$4x^2 - 16xy + 20y^2 = (x \;\; y) \begin{pmatrix} 4 & -8 \\ -8 & 20 \end{pmatrix} (x \;\; y)^T.$$

Invert the "matrix in the middle" to obtain the covariance matrix:

$$\begin{pmatrix} \sigma_X^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} 4 & -8 \\ -8 & 20 \end{pmatrix}^{-1} = \begin{pmatrix} 5/4 & 1/2 \\ 1/2 & 1/4 \end{pmatrix}$$

We conclude that

$$\sigma_X^2 = 5/4, \quad \sigma_Y^2 = 1/4, \quad \rho = \frac{2}{\sqrt{5}}.$$

Plugging these back into (30.2)-(30.3), we get the same answers we got via Method 1.

30.2 Nonlinear Transformation of Multivariate Densities

Before giving the general theory, I give an instructive example.

Example 30.1. Let X, Y be independent Gaussian(0,1) RV's. Convert the random point (X, Y) to polar coordinates (R, Θ). Find the joint PDF $f_{R,\Theta}(r, \theta)$ of R, Θ. I solve this problem using the following formula (which may be found in a calculus book) for converting double integrals from rectangular to polar coordinates:

$$\iint_{\{xy\ region\}} \phi(x, y)dxdy = \iint_{\{r\theta\ region\}} \phi(r\cos\theta, r\sin\theta)rdrd\theta \tag{30.4}$$

The function $\phi(x, y)$ can be any integrable function. The reason (30.4) is true has to do with the fact that the differential of area $dxdy$ in rectangular coordinates x, y becomes differential of area $rdrd\theta$ in polar coordinates r, θ. On the left side of (30.4), suppose that we take $\phi(x, y)$ to be $f_{X,Y}(x, y)$. Then (30.4) becomes

$$\iint_{\{xy\ region\}} f_{X,Y}(x, y)dxdy = \iint_{\{r\theta\ region\}} f_{X,Y}(r\cos\theta, r\sin\theta)rdrd\theta \tag{30.5}$$

Interpret the left side of (30.5) as a probability calculation for (X, Y) falling in some xy-region. Then we can regard the right side as giving the same answer for the probability of (R, Θ) falling in the $r\theta$-region corresponding to the xy-region in going from rectangular to polar coordinates. Since the integrand on the right side gives the right answer for any such probability calculation, that integrand must be the joint density of R, Θ. In other words, we have proved the formula

$$f_{R,\Theta}(r, \theta) = r\, f_{X,Y}(r\cos\theta, r\sin\theta).$$

In our particular case here, we have

$$f_{X,Y}(x, y) = \left(\frac{1}{2\pi}\right)\exp(-[x^2 + y^2]/2).$$

We have the following formula related polar coordinate r to rectangular coordinates x, y:

$$r^2 = x^2 + y^2.$$

Therefore,

$$f_{R,\Theta}(r, \theta) = r\left(\frac{1}{2\pi}\right)\exp(-r^2/2), \quad r > 0; \quad 0 \le \theta < 2\pi$$

We can make further conclusions about the marginal distributions of R and Θ: By the factorization rule, R and Θ must be statistically independent and their marginal PDF's are

$$f_R(r) = r\exp(-r^2/2), \quad r > 0\ (zero\ elsewhere) \tag{30.6}$$

$$f_\Theta(\theta) = \frac{1}{2\pi}, \quad 0 \le \theta < 2\pi\ (zero\ elsewhere) \tag{30.7}$$

From (30.7), we see that Θ is uniformly distributed from 0 to 2π. The density for R in (30.6) is new to us and defines what is called a *Rayleigh distribution* (see page 506 of Appendix A). The Rayleigh distribution pops up in various applications, particularly in applications to radar detection.

Discussion. Example 30.1 has pointed up one particular nonlinear transformation of interest, namely the transformation that takes you from rectangular coordinates x, y to polar coordinates r, θ. In equation (30.4), the r in the differential of area $rdrd\theta$, can be viewed as arising from the formula

$$\left| \det \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial r} \\ \frac{\partial x}{\partial \theta} & \frac{\partial y}{\partial \theta} \end{pmatrix} \right| = r. \tag{30.8}$$

The reader is invited to verify this formula by evaluating the four partial derivatives on the left side from the equations

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \end{aligned}$$

The determinant on the left side of (30.8) is called the Jacobean of the "old coordinates" x, y with respect to the "new coordinates" r, θ, and is denoted by $J(r, \theta)$. Thus, we have the following formula relating the differential of area in the old coordinates to the differential of area in the new coordinates:

$$dxdy = |J(r, \theta)|drd\theta.$$

Suppose more generally that we go from rectangular coordinates x, y to coordinates u, v in a new coordinate system, via a nonlinear transformation of the form

$$u = g_1(x, y) \tag{30.9}$$
$$v = g_2(x, y) \tag{30.10}$$

Solving for the old coordinates x, y in terms of the new coordinates u, v, we could obtain equations of the form

$$\begin{aligned} x &= h_1(u, v) \\ y &= h_2(u, v) \end{aligned}$$

The Jacobean $J(u, v)$ would be defined by

$$J(u, v) \triangleq \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{pmatrix}$$

The differential of area in x, y coordinates would be related to the differential of area in u, v coordinates by the equation

$$dxdy = |J(u, v)|dudv.$$

A double integral in the old coordinates could then be converted to a double integral in the new coordinates as follows:

$$\iint_{\{xy\ region\}} \phi(x,y)dxdy = \iint_{\{uv\ region\}} \phi(h_1(u,v),h_2(u,v))|J(u,v)|dudv \tag{30.11}$$

You can find formula (30.11) in many calculus books.[1]

Conclusion. Let X, Y be RV's with joint density $f_{X,Y}(x,y)$. Obtain new RV's U, V by transforming from x, y coordinates to u, v coordinates according to formulas (30.9)-(30.10). Then, from equation (30.11), we can conclude that the joint density $f_{U,V}(u,v)$ is obtainable via the formula

$$f_{U,V}(u,v) = f_{X,Y}(h_1(u,v),h_2(u,v))|J(u,v)| \tag{30.12}$$

Example 30.2. Let X, Y be independent continuously distributed RV's. Suppose we want to find the density of the new RV

$$U = XY.$$

(For example, X might be "current", Y might be "resistance", and U might be "voltage".) Here is how we can do this: Introduce a second spurious RV V, find $f_{U,V}(u,v)$, and then integrate out the v variable to obtain the density $f_U(u)$ of U. Let us take our "spurious" second variable V in this case as follows:

$$V = X.$$

In other words, we are performing the following nonlinear transformation from coordinates x, y to coordinates u, v:

$$\begin{aligned} u &= xy \\ v &= x \end{aligned}$$

The inverse transformation is

$$\begin{aligned} x &= v \\ y &= u/v \end{aligned}$$

and from this one determines that the absolute value of the Jacobean is

$$|J(u,v)| = \frac{1}{|v|}.$$

Plugging into (30.12), we obtain

$$f_{U,V}(u,v) = \frac{f_X(v)f_Y(u/v)}{|v|},$$

[1]For example, see page 743 of the Third Edition of *Calculus and Analytic Geometry* by Professor George Thomas (Addison-Wesley Pub. Co., 1965). This is one of the most celebrated calculus books written on Planet Earth.

and therefore we have the following formula for the PDF of $U = XY$:

$$f_U(u) = \int_{-\infty}^{\infty} \frac{f_X(x) f_Y(u/x)}{|x|} dx.$$

Exercise. Let X, Y be continuously distributed independent RV's. Let $U = X/Y$. Prove that

$$f_U(u) = \int_{-\infty}^{\infty} |y| f_X(uy) f_Y(y) dy.$$

Hint: For the transformation

$$\begin{aligned} u &= x/y \\ v &= y \end{aligned}$$

find $f_{U,V}(u, v)$ and then integrate out the variable v.

30.3 Point Estimation of a Parameter

Suppose you have a random sample $X_1, X_2, \cdots, X_n$ of size n from some prob dist with unknown parameter θ. A point estimator for θ could be any function $\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_n)$ of the random sample that does not depend on any unknown parameters. Keep in mind that your point estimator $\hat{\theta}$ is a random variable. If you have a pretty good point estimator $\hat{\theta}$, then a large percentage of the time your observation of the value of the random variable $\hat{\theta}$ will agree with θ to so many decimal places. It is our job in this section to give you some evaluation criteria that will help you decide whether a point estimator is a good point estimator. By the end of this section, you will have been exposed to three types of point estimators: (i)*unbiased estimators*, (ii) *consistent estimators*, and (iii) *minimum variance estimators*. One generally tries to select a point estimator that is simultaneously of all three types.

Examples of Point Estimators

Let μ and σ^2 be the mean and variance of the distribution you're sampling from. Then:

- The sample mean $\bar{X}_n$ is typically used as a point estimator for μ.
- If μ is known, then
$$\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n} \tag{30.13}$$
is typically used as a point estimator for σ^2.

- If μ is unknown, then the sample variance, defined earlier and given by

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n-1},$$

is typically used as a point estimator for σ^2.

30.3.1 Unbiased Point Estimators

The point estimator $\hat{\theta}(X_1, X_2, \cdots, X_n)$ is said to be *unbiased* if

$$E[\hat{\theta}(X_1, X_2, \cdots, X_n)] = \theta,$$

no matter what the value of θ is. Two reasons for seeking an unbiased estimator $\hat{\theta}$ are:

- If $\hat{\theta}$ is unbiased, its fluctuations about θ upon repeated observation tend to cancel each other out in the long run. That is, you will have

$$E[\hat{\theta} - \theta] = 0.$$

- If $\hat{\theta}$ is unbiased, then

$$Var(\hat{\theta}) = E[(\hat{\theta} - \theta)^2].$$

The number $E[(\hat{\theta} - \theta)^2]$ is the so-called "mean square estimation error". Designing the estimator to make the mean square estimation error small is therefore the same thing as making the variance of the estimator small, if you're dealing with unbiased estimators.

Examples of Unbiased Estimators

We establish that all three point estimators mentioned previously are unbiased.

- The sample mean $\bar{X}_n$ is an unbiased estimator of μ. We established this in an earlier set of notes.

- When μ is known, the estimator (30.13) is an unbiased estimator of σ^2. To establish this, note that

$$E[(X_i - \mu)^2] = \sigma^2$$

for all i. Therefore,

$$E\left[\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}\right] = \frac{n\sigma^2}{n} = \sigma^2,$$

doing the expected value term by term.

- The sample variance is an unbiased estimator of σ^2. To establish this, one first proves the identity

$$sample\ variance = \frac{n}{n-1}\left[-(\bar{X}_n - \mu)^2 + \frac{\sum_{i=1}^n (X_i - \mu)^2}{n}\right] \tag{30.14}$$

(I will prove (30.14) at the end of this subsection.) The expected value of the right side of (30.14) is

$$\frac{n}{n-1}\left[-\frac{\sigma^2}{n} + \sigma^2\right], \tag{30.15}$$

because

$$\frac{\sigma^2}{n} = Var(\bar{X}_n) = E[(\bar{X}_n - \mu_{\bar{X}_n})^2] = E[\bar{X}_n - \mu)^2].$$

Simple algebra shows that (30.15) is equal to σ^2.

Proof of (30.14). Let $x_1, x_2, \cdots, x_n$ be any real numbers and let $\bar{x}$ be the average of these numbers. I show for any constant C that

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = -(\bar{x} - C)^2 + \frac{\sum_{i=1}^n (x_i - C)^2}{n} \tag{30.16}$$

Once I prove (30.16), equation (30.14) will clearly follow. Consider the prob dist given by the density function

$$n^{-1}\sum_{i=1}^n \delta(x - x_i).$$

The mean of this prob dist is $\bar{x}$. The left side of (30.16) is therefore the variance of this prob dist. Therefore equation (30.16) is just a special case of formula (10.5) of Section 10.4 of the class notes.

30.3.2 Consistent Estimators

Suppose the point estimator $\hat{\theta}(X_1, X_2, \cdots, X_n)$ is defined no matter how big the sample size n is. We use notation $\hat{\theta}_n$ for $\hat{\theta}(X_1, X_2, \cdots, X_n)$ to denote the dependence of this estimator upon n. We say that $\hat{\theta}_n$ is a *consistent estimator* for θ if

$$\lim_{n\to\infty} P[\theta - \epsilon \le \hat{\theta}_n \le \theta + \epsilon] = 1,$$

for every $\epsilon > 0$. If you look back in Section 26.4, you will see the definition of stochastic convergence. Saying that $\hat{\theta}_n$ is a consistent estimator for θ is the same thing as saying that $\hat{\theta}_n$ converges stochastically to θ as $n \to \infty$.

The concept of consistent estimator is important for the following reason: If $\hat{\theta}_n$ is a consistent estimator of θ, then no matter how small a $\epsilon > 0$ you pick, you can pick n large enough so that

$$[\hat{\theta}_n - \epsilon,\ \ \hat{\theta}_n + \epsilon]$$

will be a confidence interval for θ having whatever preset level of confidence that you would like. For example, if you want a 90% confidence interval for θ with $\epsilon = 0.01$, you will be able to select sample size n so that

$$P[\hat{\theta}_n - 0.01 \leq \theta \leq \hat{\theta}_n + 0.01] \geq 0.90,$$

which is the statement that

$$[\hat{\theta}_n - 0.01, \ \ \hat{\theta}_n + 0.01]$$

is a 90% confidence interval for θ.

All three of the point estimators given earlier in this lecture are consistent estimators. Let's see why this is true.

- $\bar{X}_n$ is a consistent estimator for μ by the law of large numbers, established in Section 26.4.
- If μ is known, $\sum_{i=1}^{n}(X_i - \mu)^2/n$ is a consistent estimator for σ^2. To see this, note that the RV's

  $$Y_i = (X_i - \mu)^2, i = 1, 2, \cdots$$

  are independent, all have the same distribution, and therefore all have the same mean, which is σ^2. Therefore, by the law of large numbers, $(\sum_{i=1}^{n} Y_i)/n$ converges stochastically to σ^2, the mean of the Y_i's.
- It is a little bit harder to show that the sample variance (26.1) is a consistent estimator for σ^2. We use the following fact about stochastic convergecne, which is not hard to prove: if Y_n converges stochastically to a real number α and Z_n converges stochoastically to a real number β, then $A_n(Y_n + Z_n)$ converges stochastically to $\alpha + \beta$ for any sequence A_n of real numbers converging to 1. We now appeal to equation (30.15), choosing

  $$\begin{aligned} Y_n &= -(\bar{X}_n - \mu)^2 \\ Z_n &= \sum_{i=1}^{n}(X_i - \mu)^2/n \\ A_n &= \frac{n}{n-1} \end{aligned}$$

  Y_n converges stochastically to 0, Z_n converges stochastically to σ^2, and so $A_n(Y_n + Z_n)$, the sample variance, converges stochastically to $0 + \sigma^2 = \sigma^2$.

30.3.3 Minimum Variance Estimators

In this section, we restrict ourselves to all unbiased estimators of parameter θ that are based on a fixed number of samples n. Which one of these is the "best"? Let $\hat{\theta}^1$ and $\hat{\theta}^2$ be two unbiased estimators of θ, and suppose that

$$Var(\hat{\theta}^1) < Var(\hat{\theta}^2).$$

This is the same thing as saying that

$$E[(\hat{\theta}^1 - \theta)^2] < E[(\hat{\theta}^2 - \theta)^2],$$

which is the statement that the mean square estimation error for $\hat{\theta}^1$ is smaller than the mean square estimation error for $\hat{\theta}^2$. Our goal should be to select an estimator for θ with the smallest possible mean square estimation error. Suppose we can find an unbiased estimator $\hat{\theta}$ such that the variance of $\hat{\theta}$ is $\leq$ the variance of any other unbiased estimator for θ. Then we call $\hat{\theta}$ a *minimum variance* estimator. It will yield the smallest possible mean square estimation error in estimating θ.

Statisticians have devoted a lot of study towards finding minimum variance estimators. They have shown that if the prob dist you sample from is sufficiently nice, then there will exist a minimum variance estimator. In particular, if you sample from a Gaussian distribution, the following result is well known.

Result. If you sample from a Gaussian distribution with mean μ then the sample mean $\bar{X}_n$ is a minimum variance estimator for μ.

You will find this result in any good statistics book.[2] It is not that easy to prove without developing more statistical methods than we presently have at our disposal.

30.4 Random Process Introduction

In the last part of EE 3025, we consider *random processes* (also called *stochastic processes* or *random signals*). There are four kinds of random processes:

- A *discrete-time unilateral* random process is an infinite collection of random variables

  $$X_n, \quad n = n_0, n_0 + 1, n_0 + 2, \ldots$$

  where n_0 is the "starting time".

- A *discrete-time bilateral* random process is an infinite collection of random variables

  $$X_n, \quad n = 0, \pm 1, \pm 2, \pm 3, \ldots$$

- A *continuous-time unilateral* random process is an infinite collection of random variables

  $$X(t), \quad t \geq t_0$$

  where t_0 is the "starting time".

[2] For example, you may find this result in the textbook *Introduction to Mathematical Statistics: Second Edition* by Robert Hogg and Allen Craig (The MacMillan Pub. Co., 1965).

- A *continuous-time bilateral* random process is an infinite collection of random variables

$$X(t), \quad -\infty < t < \infty$$

Each time you perform an experiment on which a random process is defined, the outcome is *one entire continuous-time or discrete-time signal extending infinitely in time.* Thus, you are generating a signal at random and that is why "random signal" is a good alternate terminology for "random process". The different signals you get by performing the experiment over and over are called *realizations* of the random process. The set of all possible realizations of a random process is called the *ensemble of all realizations.* With this terminology, you can now can think of the random experiment as choosing for you a realization at random from the ensemble of realizations, on each performance of the experiment.

The realizations of discrete-time random processes are discrete-time deterministic signals, and the realizations of continuous-time random processes are continuous-time deterministic signals.

Example 30.3. A random experiment consists of flipping a fair coin *infinitely many* times. For each nonnegative integer n, define the random variable X_n to be equal to 1 if the n-th flip results in "heads", and define X_n to be -1 if the n-flip results in "tails". The discrete-time random process $X_n,\ n = 1, 2, 3, \ldots$ is called the *Bernoulli process.* The ensemble of realizations of the Bernoulli process consists of all discrete-time signals $x_n,\ n = 1, 2, 3, \ldots$, in which the signal amplitude x_n for each n is ± 1. The beginning of one particular Bernoulli process realization is plotted below. This plot tells us that the first 7 coin flips resulted in H,H,T,H,T,T,T,

Next lecture, I will give more examples of random processes.