LECTURE 20

The Gaussian (Normal) Distribution

Reading: Sections 4.6.1-4.6.3

Outline

20.0. Review

1. Continuous random variables
   - Probability density function, $f(x)dx = P(x \leq X \leq x + dx) = F'(x)dx$
   - Cumulative distribution function, $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$
   - Expected value, $E(X) = \int_{-\infty}^{\infty} xf(x)dx$
   - Variance, $\text{Var}(X) = \int_{-\infty}^{\infty}(x - E(X))^2 f(x)dx = E(X^2) - E(X)^2$.
2. General rules for expectation and variance
   - $E(aX + bY) = aE(X) + bE(Y)$
   - $\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$ if X and Y are independent.

20.1. The Normal Distribution

The most prominent distribution in statistics is the *normal*, or *Gaussian* distribution. The reason for its importance is the *Central Limit Theorem*, which will be introduced after we set out the basic mathematics of the distribution.

The *standard normal* random variable (frequently denoted Z) is described by the probability density function

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \ . \tag{20.1}$$

This is the classic "bell shaped curve".

Unit area

Because of the square in the exponent, the normal PDF function is difficult to integrate. Proving that f has unit area relies on a trick, which everyone should see once. Instead of calculating the area $A = \int f(x)dx$, we will calculate A^2.

$$\begin{aligned} A^2 &= \left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \ dz\right]^2 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \ dy \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \ dz \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-(y^2+z^2)/2} \ dydz \ . \end{aligned}$$

Change from rectangular to polar coordinates, $r = \sqrt{y^2 + z^2}$ and $dydz = rdrd\theta$. Then,

$$A^2 = \int_0^{2\pi} \int_0^{\infty} \frac{1}{2\pi} e^{-r^2/2} rdrd\theta = \int_0^{\infty} e^{-r^2/2} rdr \ .$$

Now make another change of variable, $u = r^2/2$, $du = rdr$. Then,

$$A^2 = \int_0^{\infty} e^{-u} du = 1.$$

Mean and variance

The standard normal has zero mean and unit variance.

$$E(Z) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} z e^{-z^2/2} \ dz$$

Observe that the function $ze^{-z^2/2}$ has odd symmetry. Integrating an odd function on a symmetric interval $(-a, a)$ always gives zero. Thus,

$$E(Z) = \lim_{a\to\infty} \int_{-a}^{a} \frac{1}{\sqrt{2\pi}} z e^{-z^2/2} \ dz = 0 \ .$$

For the variance, use integration by parts.

$$\mathrm{Var}(Z) = \int_{-\infty}^{\infty} (z - E(Z))^2 f(z) dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} z^2 e^{-z^2/2} dz \ .$$

With $u = z$ and $dv = ze^{-z^2/2}dz$, we have $du = dz$ and $v = -e^{-z^2/2}$. Thus,

$$\mathrm{Var}(Z) = \underbrace{-\frac{1}{\sqrt{2\pi}} z e^{-z^2/2}\bigg|_{-\infty}^{\infty}}_{0} + \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz}_{1} = 1 \ .$$

The notation $Z \sim \mathcal{N}(0, 1)$ means that Z is normally distributed with zero mean and unit standard deviation.

Cumulative distribution function

The cumulative distribution function for the standard normal r.v.,

$$F(z) = P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \ , \tag{20.2}$$

does not have a nice expression in terms of elementary functions. Numerical values are tabulated in many books (see, *e.g.*, Table 4 in the text).

The PDF $f(z)$ has even symmetry, $f(z) = f(-z)$. It follows from this that

$$F(-z) = 1 - F(z) \tag{20.3}$$

Proof

$$\begin{aligned} F(-z) &= \int_{-\infty}^{-z} f(t)dt = \int_{z}^{\infty} f(-u)du && \text{(Change variables } u = -t) \\ &= \int_{z}^{\infty} f(u)du && (f(z) = f(-z)) \\ &= P(Z > z) = 1 - P(Z \le z) = 1 - F(z) \end{aligned}$$

Two useful consequences of this are

$$F(0) = 1/2 \tag{20.4}$$

$$P(-z \le Z \le z) = F(z) - F(-z) = 2F(z) - 1 \tag{20.5}$$

The standard normal CDF is also expressed in terms of the *error function* and the *complementary error function*,

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt \tag{20.6}$$

$$\text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-t^2} dt = 1 - \text{erf}(x) \ . \tag{20.7}$$

Matlab includes the functions `erf` and `erfc`. Making the change of variable $u = t/2$,

$$F(z) = \begin{cases} \dfrac{1}{2} + \displaystyle\int_{0}^{z/\sqrt{2}} \frac{1}{\sqrt{\pi}} e^{-u^2/2} du = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{z}{\sqrt{2}}\right), & z \ge 0 \\ \\ \dfrac{1}{2} - \displaystyle\int_{0}^{-z/\sqrt{2}} \frac{1}{\sqrt{\pi}} e^{-u^2/2} du = \frac{1}{2}\text{erfc}\left(-\frac{z}{\sqrt{2}}\right), & z < 0 \end{cases} \tag{20.8}$$

General normal distribution

If a random variable Z (of any distribution) has zero mean and unit variance, then the new random variable $X = \mu + \sigma Z$ has mean μ and variance σ^2

Proof

$$E(X) = E(\mu + \sigma Z) = \mu + \sigma E(Z) = \mu$$

$$\text{Var}(X) = \text{Var}(\mu + \sigma Z) = \sigma^2 \text{Var}(Z) = \sigma^2$$

In the particular case where Z is standard *normal*, the PDF for the random variable $X = \mu + \sigma Z$ also has a normal form, with mean μ and variance σ^2 (denoted $X \sim \mathcal{N}(\mu, \sigma)$). To derive the general normal PDF, calculate the CDF $F_X(x)$, then differentiate it to get $f_X(x)$.

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(\mu + \sigma Z \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= \int_{-\infty}^{(x-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = F_Z\left(\frac{x-\mu}{\sigma}\right) . \end{aligned}$$

Now, using the chain rule,

$$\begin{aligned} f_X(x) &= \frac{d}{dx} F_X(x) = \frac{d}{dx} F_Z\left(\frac{x-\mu}{\sigma}\right) = F_Z'\left(\frac{x-\mu}{\sigma}\right) \frac{d}{dx} \frac{x-\mu}{\sigma} \\ &= f_Z\left(\frac{x-\mu}{\sigma}\right) \cdot \frac{1}{\sigma} \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} . \end{aligned} \tag{20.9}$$

Only the standard normal probabilities are tabulated. When computing a probability involving $X \sim \mathcal{N}(\mu, \sigma)$, you should standardize it first, *i.e.*, calculate $z = \dfrac{x-\mu}{\sigma}$. Then,

$$F_X(x) = P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = F_Z\left(\frac{x-\mu}{\sigma}\right) .$$

Looking up $F_Z(z)$ in a table, or computing it, will give $F_X(x)$.

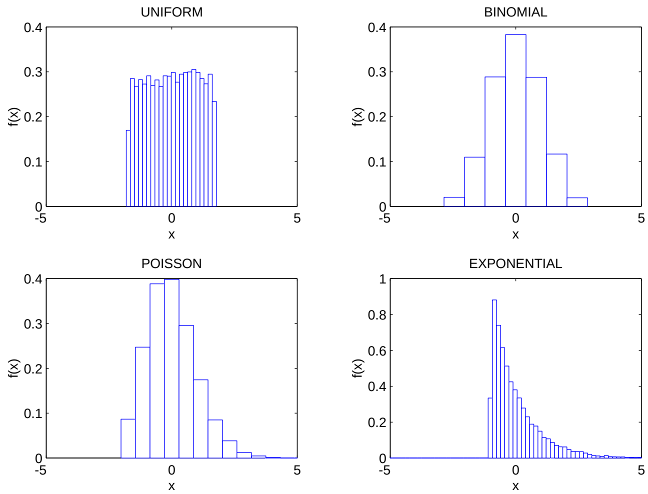
20.2. Central Limit Theorem (empirical)

Sums of random variables have a remarkable property which is summarized by the *Central Limit Theorem*:

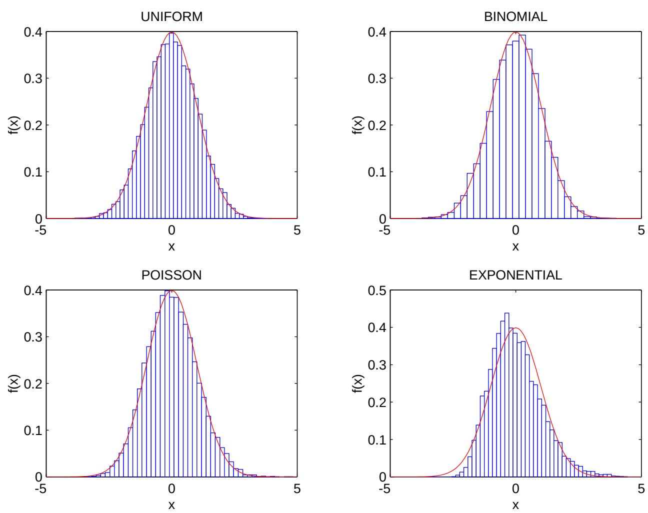
> Let $\{X_k\}_{k=1}^{n}$ be independent, identically distributed random variables with mean μ and standard deviation σ. Let the standardized random variables Y_k be defined in the usual way, $Y_k = \dfrac{X_k - \mu}{\sigma}$, and form the sum
>
> $$Y = \sum_{k=1}^{n} \frac{Y_k}{\sqrt{n}} .$$
>
> In the limit as $n \to \infty$, Y is a standard normal random variable.

The proof is interesting, but a picture is worth a thousand words. Shown below are two sets of graphs resulting from a simulation experiment. From each of four distributions — binomial, Poisson, uniform, and exponential — a sample of 10000 random variates X was computed and standardized. Their empirical distributions are displayed in the first set of graphs. Then, for each distribution, nine additional sets of 10000 random variates were computed and standardized, and 10000 sums Y of ten variates each were calculated. Their empirical distributions are displayed in the second set of graphs. Even with only $n = 10$, the approach to a standard normal distribution (smooth curve) is observed.

n = 1



n = 10



Electronic noise is often modeled as a Gaussian random process, *i.e.*, as a time function $\nu(t)$ which, at every value of t, is a normally-distributed random variable. Such a model agrees with experiment. It also makes physical sense. Noise currents are the result of the random movement of huge numbers of electrons. By the central limit theorem, the sum of these random currents is normally distributed.

20.3. Normal approximations to the binomial and Poisson distributions

A Bernoulli random variable X with probability of success p has mean p and variance $p(1-p)$. According to the central limit theorem then, the random variable

$$Y = \sum_{k=1}^{n} \frac{X_k - p}{\sqrt{n}\sqrt{p(1-p)}} = \frac{\sum_{k=1}^{n}(X_k - p)}{\sqrt{np(1-p)}} = \frac{\left(\sum_{k=1}^{n} X_k\right) - np}{\sqrt{np(1-p)}}$$

is approximately $\mathcal{N}(0,1)$ for large n. But the sum $\sum_{k=1}^{n} X_k$ is the sum of n identical, independent Bernoulli trials, *i.e.*, a binomial random variable with mean np and variance $np(1-p)$. Thus, the central limit theorem says that for sufficiently large n, the binomial PDF is approximately normal with $\mu = np$ and $\sigma^2 = np(1-p)$,

$$\binom{n}{x} p^x (1-p)^{n-x} \approx \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left(-\frac{(x-np)^2}{2np(1-p)}\right) \ . \tag{20.10}$$

When n is large and p is small, the binomial distribution is approximately Poisson with $\mu = np$. This implies that, for large μ, the Poisson distribution is approximately normal, with mean and variance μ.

$$\frac{e^{-\mu}\mu^x}{x!} \approx \frac{1}{\sqrt{2\pi\mu}} \exp\left(-\frac{(x-\mu)^2}{2\mu}\right) \ . \tag{20.11}$$

Correction for continuity

The normal distribution is often easier to work with than the binomial or Poisson. However, when using a continuous distribution to approximate a discrete distribution, the probability $P(X = a)$ should be approximated by integrating the continuous PDF over the interval $(a - \frac{1}{2}, a + \frac{1}{2})$:

$$P(X = a) \approx \int_{a-1/2}^{a+1/2} f(x)dx = F(a + 1/2) - F(a - 1/2) \tag{20.12}$$

Moreover,

$$P(X \le a) \approx \int_{-\infty}^{a+1/2} f(x)dx = F(a + 1/2) \ . \tag{20.13}$$

This is called the *correction for continuity* (M. DeGroot, *Probability and Statistics*, pp. 283-285). For example, with X a binomial(20, 1/2) rv, $P(X \le 10) = 0.588$ (text, Table 2). Approximating the binomial distribution with a normal having $\mu = 10$ and $\sigma^2 = 5$, the approximate probability without correction would be (text, Table 4)

$$P(X \le 10) = P\left(Z \le \frac{10 - 10}{2.24}\right) = P(Z \le 0) = 0.5 \ .$$

On the other hand, using the correction,

$$P(X \le 10) = P\left(Z \le \frac{10.5 - 10}{2.24}\right) = P(Z \le 0.22) = 0.5871 \ .$$