Normal Distributions and Sample Statistics

18.650, Sept. 9, 2015

1 Review of some probability

Recall that for a real-valued random variable X, its *distribution function* is the function $F(x) := F_X(x) = \Pr(X \leq x)$. Distribution functions are characterized by the properties of being nondecreasing, continuous from the right, and satisfying $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$. A probability density is a function $f(x) \geq 0$ such that $\int_{-\infty}^{\infty} f(x)dx = 1$. It is related to a distribution function F by $F(x) \equiv \int_{-\infty}^{x} f(u)du$, and $F'(x) = f(x)$ except possibly for x in a set A over which $\int_A f(x)dx = 0$.

Recall that $\exp(y)$ is a notation for e^y. A basic normal distribution is the *standard normal* distribution which has the *standard normal density* ϕ given by $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$, $-\infty < x < \infty$. From it we get the *standard normal distribution function* Φ, $\Phi(x) = \int_{-\infty}^{x} \phi(u)du$. It is tabulated in Rice, Appendix B, Table 2, p. A7, for $0 \leq x \leq 3.49$. For $x > 0$, $\Phi(-x)$ can be found as $1 - \Phi(x)$, because $\phi(-x) \equiv \phi(x)$, so

$$\Phi(-x) = \int_{-\infty}^{-x} \phi(u)du = \int_{x}^{+\infty} \phi(v)dv = 1 - \Phi(x).$$

The probability $\Phi(-3.5) \doteq 0.0002363$, which is quite small.

There are families of probability distributions depending on what are called parameters. Two examples for discrete distributions are:
(i) Binomial distributions: let X be the number of successes in n independent trials with probability p of success on each trial. Then $\Pr(X = k) = \binom{n}{k}p^k(1-p)^{n-k}$ for $k = 0, 1, ..., n$. For a fixed n, p is the parameter.
(ii) Poisson distributions, having one parameter λ, with $0 \leq \lambda < +\infty$, and $\Pr(X = k) = e^{-\lambda}\lambda^k/k!$ for $k = 0, 1, 2, ...$ (0^0 is defined as 1). This distribution is the limit of binomial distributions where $n \to \infty$, $p \to 0$, and $np \to \lambda$.

For any probability density f on the real line, one can create what is called a location-scale family as follows. For any $\sigma > 0$ and any real μ,

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$

always defines a probability density. If the distribution with density f has mean 0 and variance 1, then the one with density $f_{\mu,\sigma}$ will have mean μ and variance σ^2, thus standard deviation σ, as one can see from the change of variables $z = (x-\mu)/\sigma$, $x = \sigma z + \mu$ in integrals. Here μ is called a location parameter and σ a scale parameter.

By far the most important example of a location-scale family on the line is the family of normal distributions, which one gets starting with f equal to the standard normal density ϕ. The normal distribution with mean μ and variance σ^2, with notation $N(\mu, \sigma^2)$, has a density defined for any real μ and $0 < \sigma < \infty$ by

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty.$$

Such distributions are sometimes called Gaussian by probabilists. For reasons that statisticians don't call them that, see the historical notes in normalappend.pdf on the course website.

Normal distributions arise as limits in the central limit theorem. Variables $X_1, ..., X_n$ are called i.i.d. (independent and identically distributed) if they are jointly independent and all have the same distribution. If X_j are i.i.d. with a distribution having a finite variance $\sigma^2 > 0$ (and consequently a finite mean μ), and $S_n := X_1 + \cdots + X_n$ then the central limit theorem (given in many probability texts) says that as $n \to \infty$, $(S_n - n\mu)/(\sqrt{n}\sigma)$ converges in distribution to $N(0,1)$, in other words for all x,

$$\Pr\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \le x\right) \to \Phi(x)$$

as $n \to \infty$. Informally, one can say that S_n itself has approximately a $N(n\mu, n\sigma^2)$ distribution.

There is a more general form of the central limit theorem saying that the sum of small, independent random variables is approximately normal (Lindeberg's Theorem, normalappend.pdf). For example, errors in physical measurements have been thought to be normal because they may result from

small, independent contributions. Here for a sum $S_i = \sum_{j=1}^{n_i} X_{ij}$ and a large n_i, for X_{ij} to be "small" means small in relation to the sum S_i: no individual term has a major effect on the sum. But, as is seen in problem set 1, measurements of the same physical quantity may not be normal, even though repeated measurements with the same method and apparatus might be, but measurements in different laboratories, and/or by different methods, could have different means and/or variances. Even repeated measurements by one method in one lab might not be (exactly) normal or independent.

If $X_1, ..., X_n$ are i.i.d. each with a $N(\mu, \sigma^2)$ density, then their joint n-dimensional density is

$$\frac{1}{(\sigma\sqrt{2\pi})^n} \prod_{j=1}^{n} \exp\left(-\frac{(x_j - \mu)^2}{2\sigma^2}\right) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \mu)^2\right). \quad (1)$$

2 Statistics

In statistics, one typically begins with some data (observations) $X_1, ..., X_n$. Suppose they are real-valued. One doesn't necessarily have a good reason to assume they are i.i.d. normal.

A text file "rsystem" on the R computer language and system for statistics is posted on the course website. One can test whether a given data set $x = (X_1, ..., X_n)$ are i.i.d. $N(\mu, \sigma^2)$ for some unspecified μ and σ by the Shapiro–Wilk test, which is implemented in the R system by "shapiro.test(x)". The test will issue a "p-value," and if that is less than the conventional level 0.05, one would reject the hypothesis of normality. It might not then be justified to assume that the data are normal. If the hypothesis is not rejected, that doesn't mean it is proved or "accepted." For example, recall that the $U[0, 1]$ (uniform) distribution has density $f(x) = 1$ for $0 \le x \le 1$ and 0 elsewhere. Suppose as in PS1, one generates 25 variables i.i.d. $U[0, 1]$ by the command x = runif(25) and then tests the data vector for normality by shapiro.test(x), giving a p-value larger than .05, so that normality is not rejected. But it certainly is not accepted, because we know the sample was generated as $U[0, 1]$, not normal. Rather, we can say that for $n = 25$ the Shapiro–Wilk test is not powerful enough to reject normality for the $U[0, 1]$ sample. For a large enough n, it would be rejected.

For a sample of real data, non-rejection of normality means one is justified for the time being in going ahead and using methods based on assuming normality.

If the X_j are i.i.d. normal, then, as usually in statistics, μ and σ are unknown, but they can be estimated from the data. The mean μ can be estimated by the *sample mean* $\overline{X} = S_n/n$ where $S_n = \sum_{j=1}^n X_j$. The variance σ^2 can be estimated for $n \geq 2$ by

$$S^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X})^2.$$

Rice uses the notation S^2 for this. It is sometimes called s_x^2 (some calculators display s_x) or s_X^2. S, or s_X, is called the *sample standard deviation* for the sample $X = (X_1, ..., X_n)$. The factor $1/(n-1)$ is used, for one reason, because if X_j are i.i.d. with any distribution having a variance σ^2 with $0 < \sigma < \infty$, then $Es_X^2 = \sigma^2$ (the expectation of the sample variance equals the true variance), whereas

$$E\frac{1}{n} \sum_{j=1}^{n} (X_j - \overline{X})^2 = \frac{n-1}{n} \sigma^2.$$

But for some purposes, a factor $1/n$ rather than $1/(n-1)$ will be used.

The sample mean $\overline{X}$ is an approximation to the true mean μ which tends to get better as n increases. To see this quantitatively we need to look at the distribution of $\overline{X}$ for given μ, σ, and n.

3 More probability

The notation $X \sim P$ means the random variable X has the probability distribution P. The sum of two independent normal variables, with any means and variances, is normal:

Theorem 1 *If X and Y are independent random variables with normal distributions, $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\nu, \tau^2)$ then $X + Y$ is also normal, with $X + Y \sim N(\mu + \nu, \sigma^2 + \tau^2)$.*

This is proved in the "addnormals.pdf" handout posted on the course website. Paper copies aren't being distributed because we assume many of you know this fact from a probability course.

The next fact is stated early in Section 6.3 of Rice, p. 195.

Theorem 2 *Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Then $\overline{X}$ has the distribution $N(\mu, \sigma^2/n)$.*

Proof. For any distribution F having finite mean μ and variance σ^2, if $X_1, \ldots, X_n$ are i.i.d. (F), then $\overline{X}$ has mean μ and variance σ^2/n. So the only problem is to show that $\overline{X}$ has a normal distribution in this case. Now, S_n defined as $X_1 + \cdots + X_n$ has a normal distribution, specifically $N(n\mu, n\sigma^2)$, by Theorem 1 and induction. Multiplying by a constant $1/n$ gives $\overline{X}$ which then has the stated distribution, Q.E.D.

4 χ^2 distributions

To describe the distribution of the random variable $S^2 = s_X^2$, and for other purposes in statistics, we need the notion of a chi-squared (χ^2) distribution. If $Z_1, ..., Z_d$ are i.i.d., each having a $N(0,1)$ distribution, then $Z_1^2 + \cdots + Z_d^2$ is said to have a $\chi^2(d)$ distribution, or a chi-squared distribution with d degrees of freedom.

Next, we have a theorem that includes Corollary A and Theorem B in Section 6.3 of Rice. It gives the distribution of s_X^2 (depending on σ^2) and its independence of $\overline{X}$ in the normal case.

Theorem 3 *If $X_1, ..., X_n$ are i.i.d. $N(\mu, \sigma^2)$, $n \geq 2$, then*
(a) $\overline{X}$ and s_X^2 are independent random variables;
(b) $(n-1)s_X^2/\sigma^2$ has a $\chi^2(n-1)$ distribution.

Proof. Let $Y_j = X_j - \mu$ for $j = 1, \ldots, n$. Then $Y_1, ..., Y_n$ are i.i.d. $N(0, \sigma^2)$, $\overline{Y} = \overline{X} - \mu$ and $s_Y^2 = s_X^2$. So we can assume $\mu = 0$.

It's convenient to make a rotation of coordinates in n-space. Let the standard basis vectors be $\delta_i = \{\delta_{ij}\}_{j=1}^n$ where $\delta_{ij} = 1$ for $i = j$ and 0 for $i \neq j$. The first element of the new basis will be $e_1 = (1/\sqrt{n}, \ldots, 1/\sqrt{n})$. This does have length 1. Then we can always find further orthonormal basis vectors $e_2, ..., e_n$, for example
$e_2 = (1/\sqrt{2}, -1/\sqrt{2}, 0, \ldots, 0)$, $e_3 = (1/\sqrt{6}, 1/\sqrt{6}, -2/\sqrt{6}, 0, \ldots, 0)$,...

For any two vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ (with respect to the standard basis) we have the usual dot product $x \cdot y = \sum_{j=1}^n x_j y_j$, with the squared length of x given by $|x|^2 = x \cdot x$.

Now, for the random vector $X = (X_1, \ldots, X_n)$ we have $\overline{X} = X \cdot e_1/\sqrt{n}$, and $(\overline{X}, \ldots, \overline{X}) = (X \cdot e_1)e_1$, which is the projection of X to the e_1 axis. The lengths of vectors and their dot products are preserved by rotations of

coordinates, so

$$\sum_{j=1}^{n}(X_j - \overline{X})^2 \;=\; |X - (X\cdot e_1)e_1|^2 \;=\; \sum_{i=2}^{n}(X\cdot e_i)^2.$$

Since $X_1, \ldots, X_n$ are i.i.d. $N(0,\sigma^2)$, their joint density is by (1)

$$(\sigma\sqrt{2\pi})^{-n}\Pi_{j=1}^{n}\exp(-x_j^2/(2\sigma^2)) \;=\; (\sigma\sqrt{2\pi})^{-n}\exp(-|x|^2/(2\sigma^2)).$$

This distribution is invariant under any rotation of coordinates (change of orthonormal basis), specifically $|x|^2 = (x\cdot e_1)^2 + (x\cdot e_2)^2 + \cdots + (x\cdot e_n)^2$. Thus $X\cdot e_1,\ \ldots, X\cdot e_n$ are i.i.d. $N(0,\sigma^2)$ and $X\cdot e_i/\sigma$ are i.i.d. $N(0,1)$. It follows that $\overline{X} = X\cdot e_1/\sqrt{n}$ is independent of $s_X^2 = (n-1)^{-1}\sum_{i=2}^{n}(X\cdot e_i)^2$, proving (a). Also, $(n-1)s_X^2/\sigma^2 = \sum_{i=2}^{n}(X\cdot e_i)^2/\sigma^2$ has a $\chi^2(n-1)$ distribution, proving (b), Q.E.D.

Here is another way of looking at chi-squared distributions. As noted in the above proof, if $X_1, ..., X_d$ are i.i.d. $N(0,1)$, their joint density is $(2\pi)^{-d/2}\exp(-|x|^2/2)$ on d-dimensional space. Let $Y = X_1^2 + \cdots + X_d^2$, so that Y has a $\chi^2(d)$ distribution. We have $P(Y \le t) = 0$ for $t \le 0$. For $t > 0$, $P(|Y| \le t)$ is the integral of the density over the region where $|x|^2 \le t$, or equivalently $|x| \le \sqrt{t}$. Suppose $d \ge 2$. Using spherical coordinates, the integral becomes $A_d(2\pi)^{-d/2}\int_0^{\sqrt{t}} r^{d-1}\exp(-r^2/2)dr$ where A_d is a constant depending on d, the $(d-1)$-dimensional surface area of the unit sphere $|x| = 1$ in d-space. By the substitution $x = r^2$, $r = \sqrt{x}$, $dr = dx/(2\sqrt{x})$, the integral becomes

$$A_d(2\pi)^{-d/2}\int_0^t x^{(d-2)/2}\exp(-x/2)dx/2.$$

Since $(d-2)/2 = (d/2)-1$, and a probability density has a unique normalizing constant, this gives a proof that the $\chi^2(d)$ distribution is the $\Gamma(d/2, 1/2)$ distribution, as proved by another method in gammabeta.pdf, Theorem 4. Moreover, since we know that the normalizing constant is $(1/2)^{d/2}/\Gamma(d/2)$, we can evaluate $A_d = 2\pi^{d/2}/\Gamma(d/2)$. For example, if $d = 2$, since $\Gamma(1) = 0! = 1$, we get $A_2 = 2\pi$, the circumference of the unit circle as desired. If $d = 3$, then by the recursion formula, $\Gamma(3/2) = \Gamma(1/2)/2 = \sqrt{\pi}/2$, so $A_3 = 4\pi$, which is in fact the area of the unit sphere in 3 dimensions. Also, the volume of the unit ball $\{|x| \le 1\}$ in d dimensions is $V_d = A_d\int_0^1 r^{d-1}dr = A_d/d = \pi^{d/2}/\Gamma((d/2)+1)$, giving $V_2 = \pi$ and $V_3 = 4\pi/3$ as desired.

There are an appendix and notes in a separate file, normalappend.pdf, on the course website.