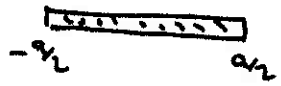
Michael E. Peskin
July 11-12, 2011

Basic Training in Statistics - Appendices

Appendix A: Useful information about distributions used in these lectures

Box distribution

$$p(x) = \begin{cases} \frac{1}{a} & -\frac{a}{2} < x < \frac{a}{2} \\ 0 & \text{otherwise} \end{cases}$$



$$\langle x \rangle = 0 \qquad \text{Var}(x) = \frac{a^2}{12}$$

Binomial distribution

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\langle k \rangle = np \qquad \text{Var}(k) = np(1-p)$$

Exponential distribution

$$p(t) = \frac{1}{\tau} e^{-t/\tau}$$

$$\langle t \rangle = \tau \qquad \text{Var}(t) = \tau^2$$

Poisson distribution

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$\langle k \rangle = \lambda \qquad \text{Var}(k) = \lambda$$

Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

$$\langle x \rangle = \mu \qquad \mathrm{Var}(x) = \sigma^2$$

χ^2 distribution

$$p(z) = \frac{1}{2^{N/2}\Gamma(N/2)} z^{N/2-1} e^{-z/2}$$

$$\langle z \rangle = n \qquad \mathrm{Var}(z) = 2n$$

Student's t distribution

$$p(t) = \frac{1}{\sqrt{N\pi}} \frac{\Gamma\left(\frac{N+1}{2}\right)}{\Gamma(N/2)} \left(1 + t^2/N\right)^{-\frac{N+1}{2}}$$

$$\langle t \rangle = 0 \qquad \mathrm{Var}(t) = \frac{N}{N-2}$$

Appendix B: Derivation of Students's t distribution

Let z follow a χ^2 distribution with N degrees of freedom. Then

$$p(z) = \frac{1}{2^{N/2}\Gamma(N/2)} z^{N/2-1} e^{-z/2}$$

Let x follow a Gaussian distribution with mean 0 and variance 1. Then

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

The joint probability distribution for z and x is the product of these factors.

Let

$$t = \frac{x}{\sqrt{z/N}}$$

If we ignore the separate values of z and x, we can integrate over these subject to the constraint that the value of t be fixed. Then the pdf for t will be

$$p(t) = \int_{-\infty}^{\infty} dx \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \int_0^{\infty} dz \frac{1}{2^{N/2}\Gamma(N/2)} z^{N/2-1} e^{-z/2} \delta\left(t - \frac{x}{\sqrt{z/N}}\right)$$

Integrate the δ function over z, using

$$\int dz \; \delta\left(t - \frac{x}{\sqrt{z/N}}\right) = \frac{2z}{|t|} \quad \text{with} \quad z = n\left(\frac{x}{t}\right)^2$$

This gives

$$p(t) = \int_{-\infty}^{\infty} dx \; \frac{2}{\sqrt{2\pi}} \; \frac{1}{2^{N/2}\Gamma(N/2)} \; \frac{N^{N/2} x^N}{|t|^{N+1}} \; e^{-\frac{1}{2}x^2(1+\frac{N}{t^2})}$$

Now change variables to

$$y = \tfrac{1}{2}x^2(1+\tfrac{N}{t^2})$$

Be careful that there are two solutions x for each y. The integral becomes

$$p(t) = \int_0^{\infty} dy \; \frac{1}{\sqrt{\pi}} \; \frac{1}{\Gamma(N/2)} \; \frac{1}{(1+N/t^2)^{\frac{N+1}{2}}} \; \frac{N^{N/2}}{(t^2)^{\frac{N+1}{2}}} \; y^{\frac{N-1}{2}} e^{-y}$$

This is a Gamma function, and so

$$p(t) = \frac{1}{\sqrt{N\pi}} \; \frac{\Gamma(\frac{N+1}{2})}{\Gamma(N/2)} \; (1+t^2/N)^{-\frac{N+1}{2}}$$

Appendix C: Derivation of the Kolmogorov-Smirnov distribution

Consider first the case of testing whether a set of points $\{x_i\}$, $i = 1, \ldots, N$, could be drawn from a probability distribution $p(x)$. Let $P(x)$ be the lower cumulative distribution associated with $p(x)$, and let $S(x)$ be the lower cumulative distribution of the data, which is a step function with values k/N. The Kolmogorov-Smirnov test requires the probability distribution of D, defined by

$$D = \max_x |S(x) - P(x)|$$

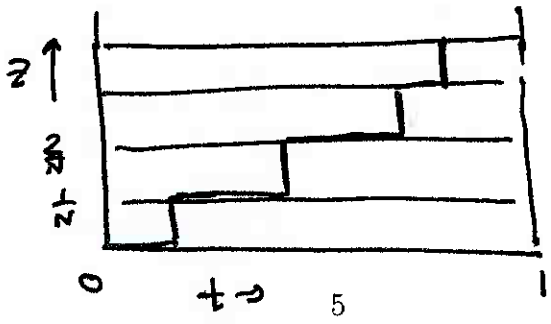
First, notice that the test is invariant with respect to reparametrization of x. Any variables $y(x)$ will give the same test. We can use this freedom to choose the variable as

$$z = P(x)$$

so that

$$P(z) = z \qquad p(z) = 1 \qquad 0 \le z \le 1$$

The realizations of $\{z_i\}$ are generated from this distribution by the following stochastic process: Let t be a variable that runs from 0 to 1. Consider a random walker who begins at $z = 0$ at $t = 0$. In any interval dt, the walker may jump forward by a distance $1/N$ with probability Ndt. We are specifically interested in walks that end at $z = 1$ at $t = 1$. Then the trajectory of the walk will be a realization of $S(z)$.



The probability distribution of z at time t obeys the equation

$$\frac{d}{dt} P(m,t) = N\left[P(m-1,t) - P(m,t)\right] \qquad z = \frac{m}{N}$$

We could carry out the analysis of this equation for any finite value N, but now I will specialize to the case $N \gg 1$. For this limit, it makes sense to consider the probability as a continuous variable of z and take the continuum limit of the equation,

$$\frac{\partial}{\partial t} P(z,t) = N\left[P(z-\tfrac{1}{N},t) - P(z,t)\right]$$

$$= N\left(-\frac{1}{N}\frac{\partial}{\partial z}P + \frac{1}{2N^2}\frac{\partial^2}{\partial z^2}P + \cdots\right)$$

Actually, we are interested in the probability distribution of

$$D = z - t$$

The probability distribution of D obeys

$$\frac{\partial}{\partial t} P(D,t) = \left(-\frac{\partial}{\partial D}P + \frac{1}{2N}\frac{\partial^2}{\partial D^2}P + \cdots\right) + \frac{\partial}{\partial D}P$$

$$\frac{\partial}{\partial t}P = \frac{1}{2N}\frac{\partial^2}{\partial D^2}P$$

This is a simple diffusion equation in 1 dimension.

We are interested in computing $P(D_*)$, the probability that the value of D is never as large as D_* at any point on the interval $0 \le t \le 1$. To address this problem, modify the original problem by placing absorbing walls at $D = \pm D_*$. Then

$$P(D_*) = \frac{P(D=0,\, t=1)\big|_{\text{walls}}}{P(D=0,\, t=1)}$$

Both diffusion equations are easily solved. The solution of the original equation is

$$P(D,t) = \frac{1}{\sqrt{2\pi t/N}}\, e^{-\frac{N}{2t}D^2}$$

The problem with absorbing walls is solved by the method of images

$$P(D,t)\Big|_{\text{walls}} = \sum_{j=-\infty}^{\infty} (-1)^j \frac{1}{\sqrt{2\pi t/N}}\, e^{-\frac{N}{2t}(D-2jD_*)^2}$$

The ratio of these functions gives the Kolmogorov-Smirnov distribution

$$P(D_*) = \sum_{j=-\infty}^{\infty} (-1)^j\, e^{-2N D_*^2}$$

Finally, consider the problem of testing whether two sets of points $\{x_{1i}\}$, $i = 1,\ldots,N_1$, and $\{x_{2i}\}$, $i = 1,\ldots,N_2$, could be generated by the same distribution. Again, we can choose a variable such the the underlying distribution is uniform. The realizations of the difference between the cumulative distributions of x_{1i} and x_{2i} is generated by a walker that, in each interval of time, steps in the positive direction by $1/N_1$ with probability $N_1 dt$ and steps in the negative direction by $/N_2$ with probability $N_2 dt$. The exact equation for the probability of D is then

$$\frac{\partial}{\partial t}P(D,t) = N_1\left[P\left(D-\tfrac{1}{N_1},t\right) - P(D,t)\right] + N_2\left[P\left(D+\tfrac{1}{N_2},t\right) - P(D,t)\right]$$

The continuum approximation, valid for large N_1, N_2 is

$$\frac{\partial}{\partial t}P(D,t) = \left(\frac{1}{2N_1}+\frac{1}{2N_2}\right)\frac{\partial^2}{\partial D^2}P(D,t)$$

Thus, this problem reduces to the previous one with the identification

$$\frac{1}{N} = \frac{1}{N_1} + \frac{1}{N_2}$$