Statistics Formulae

Mean and Variance

A random variable X has a distribution over some subset x of the real numbers. When the distribution of X is discrete, the probability that $X = x_i$ is P_i. When the distribution is continuous, the probability that X lies in an interval δx is $f(x)\delta x$, where $f(x)$ is the probability density function.

$$\text{Mean } \mu = E(X) = \sum P_i x_i \text{ or } \int x\, f(x)\, \mathrm{d}x.$$

$$\text{Variance } \sigma^2 = V(X) = E[(X-\mu)^2] = \sum P_i (x_i - \mu)^2 \text{ or } \int (x-\mu)^2 f(x)\, \mathrm{d}x.$$

Probability distributions

Error function: $\operatorname{erf}(x) = \dfrac{2}{\sqrt{\pi}} \displaystyle\int_0^x \mathrm{e}^{-y^2}\, \mathrm{d}y$

Binomial: $f(x) = \dbinom{n}{x} p^x q^{n-x}$ where $q = (1-p), \quad \mu = np, \sigma^2 = npq, p < 1.$

Poisson: $f(x) = \dfrac{\mu^x}{x!}\mathrm{e}^{-\mu}$, and $\sigma^2 = \mu$

Normal: $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left[-\dfrac{(x-\mu)^2}{2\sigma^2}\right]$

Weighted sums of random variables

If $W = aX + bY$ then $E(W) = aE(X) + bE(Y)$. If X and Y are independent then $V(W) = a^2V(X) + b^2V(Y)$.

Statistics of a data sample $x_1, \ldots, x_n$

$$\text{Sample mean } \overline{x} = \frac{1}{n}\sum x_i$$

$$\text{Sample variance } s^2 = \frac{1}{n}\sum (x_i - \overline{x})^2 = \left(\frac{1}{n}\sum x_i^2\right) - \overline{x}^2 = E(x^2) - [E(x)]^2$$

Regression (least squares fitting)

To fit a straight line by least squares to n pairs of points (x_i, y_i), model the observations by $y_i = \alpha + \beta(x_i - \overline{x}) + \epsilon_i$, where the ϵ_i are independent samples of a random variable with zero mean and variance σ^2.

Sample statistics: $s_x^2 = \dfrac{1}{n}\sum (x_i - \overline{x})^2, \quad s_y^2 = \dfrac{1}{n}\sum (y_i - \overline{y})^2, \quad s_{xy}^2 = \dfrac{1}{n}\sum (x_i - \overline{x})(y_i - \overline{y}).$

Estimators: $\widehat{\alpha} = \overline{y}, \widehat{\beta} = \dfrac{s_{xy}^2}{s_x^2}; E(Y \text{ at } x) = \widehat{\alpha} + \widehat{\beta}(x - \overline{x}); \widehat{\sigma}^2 = \dfrac{n}{n-2}(\text{residual variance}),$

where residual variance $= \dfrac{1}{n}\sum \{y_i - \widehat{\alpha} - \widehat{\beta}(x_i - \overline{x})\}^2 = s_y^2 - \dfrac{s_{xy}^4}{s_x^2}.$

Estimates for the variances of $\widehat{\alpha}$ and $\widehat{\beta}$ are $\dfrac{\widehat{\sigma}^2}{n}$ and $\dfrac{\widehat{\sigma}^2}{ns_x^2}$.

Correlation coefficient: $\widehat{\rho} = r = \frac{s_{xy}^2}{s_x s_y}$.