# 17 Paired t-test

Sometimes we have to deal with data that come in linked pairs. For example initial weight and final weight for a grwoth harmone. We have observations $(x_i, y_i)$. They are very higly correlated. So if we want to test that the harmone has no effect, or equivalently that means of $\{x_i\}$ and $\{y_i\}$ are the same, we cannot afford to assume that $\{x_i\}$ and $\{y_i\}$ are independent. On the other hand we can form the differences $d_i = x_i - y_i$ and test for the mean of $\{d_i\}$ to be 0. This is now a simple $t$ test.

# 18 Correlation and Regression

Often we have two related variables $X$ and $Y$ and only one of them say $X$ is observed. We would like to use the observed value of $X$ to predict $Y$. If we know the joint distribution of $(X, Y)$, we can determine the conditional distribution of $Y$ given $X$ and use it as a guide to our prediction. The expectation of the conditional distribution or the conditional expectation is a reasonable guess regaridng what we might expect for $Y$.

The conditonal expectation has the property that it minimizes $E[(Y - f(X))^2]$ over all functions $f(X)$ that depend only on $X$.

In practice one often mnimizes $E[(Y - f(X))^2]$ over a specified limited class of functions $f(\cdot)$. In linear regreesion one limits the choice of $f$ to linear functions of $X$ of the form $a + bX$ where $a$ and $b$ are constants.

The minimization

$$\inf_{a,b} E\big[(Y - a - bX)^2\big]$$

can be explicitly carried out. Differentiating with repect to $a$ and $b$ we get the equations

$$\begin{aligned} E\big[(Y - a - bX)\big] &= 0 \\ E\big[(Y - a - bX)X\big] &= 0 \end{aligned}$$

or

$$\begin{aligned} E\big[Y\big] &= a + bE\big[X\big] \\ E\big[XY\big] &= aE\big[X^2\big] + bE\big[X\big] \end{aligned}$$

Solving the equations for $a$ and $b$ we get

$$\hat{b} = \frac{E\big[XY\big] - E\big[X\big]E\big[Y\big]}{E\big[X^2\big] - (E\big[X\big])^2} = \frac{\text{Cov } XY}{\text{Var } X}$$
$$\hat{a} = E\big[Y\big] - \hat{b}E\big[X\big]$$

We can therefore write the regression line as

$$Y - E\big[Y\big] = \frac{\text{Cov } XY}{\text{Var } X}\big[X - E\big[X\big]\big]$$

For any two random variables $X$ and $Y$, the covariance is defined by Cov $[XY] = E[XY] - E[X]E[Y]$. Note that Cov $[XX] = E[X^2] - (E[X])^2] =$ Var $[X]$.

One can decompose $Y - E[Y]$ as

$$Y - E[Y] = \hat{b}\big[X - E[X]\big] + Y - \hat{a} - \hat{b}X$$

Because the cross term vanishes, we get

$$\text{Var } \big[Y\big] = \frac{\big(\text{Cov } \big[XY\big]\big)^2}{\text{Var}\big[X\big]\ \text{Var}\big[Y\big]}\text{Var } \big[Y\big] + \left[1 - \frac{\big(\text{Cov } \big[XY\big]\big)^2}{\text{Var}\big[X\big]\ \text{Var}\big[Y\big]}\right]\text{Var } \big[Y\big]$$

The linear correlation coefficient $\rho$ between $X$ and $Y$ is defined as

$$\rho = \frac{\text{Cov } \big[XY\big]}{\sqrt{\text{Var } \big[X\big]}\sqrt{\text{Var } \big[Y\big]}}$$

We can rewrite the earlier relation as

$$\text{Var } \big[Y\big] = \rho^2\text{Var } \big[Y\big] + (1 - \rho^2)\text{Var } \big[Y\big]$$

The first term is the amount of reduction in the variance due to "predictable" and the second term is the residual variance. $\rho^2$ is the proportion of the variance that is reduced and $(1-\rho^2)$ is the residual proportion. From Schwartz's inequality it is clear that $-1 \leq \rho \leq 1$.

If we have observations $(x_i, y_i)$ from a bivariate population, we can estimate the means, variances and covariances by their corresponding sample

values

$$\bar{x} = \frac{1}{n}\sum_i x_i$$
$$\bar{y} = \frac{1}{n}\sum_i y_i$$
$$s_x^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2$$
$$s_y^2 = \frac{1}{n}\sum_i y_i^2 - \bar{y}^2$$
$$c_{x,y} = \frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}$$
$$\hat{b} = \frac{c_{x,y}}{s_x^2}$$
$$r = \frac{c_{x,y}}{s_x s_y}$$

These are clearly consistent estimators of the corresponding population values.

We can interchange the roles of $X$ and $Y$ and there is the regression line

$$X - E\big[X\big] = \frac{\text{Cov } XY}{\text{Var } Y}\big(Y - E\big[Y\big]\big)$$

While both lines pass through $\big(E\big[X\big], E\big[Y\big]\big)$, they have in general different slopes unless

$$\frac{\text{Cov } \big[XY\big]}{\text{Var } \big[X\big]} = \frac{\text{Var } \big[Y\big]}{\text{Cov } \big[XY\big]}$$

or

$$\rho^2 = 1$$

whch corresponds to an exact linear relation between $X$ and $Y$.

# 19 Multivariate Normal Distributions

Just as the family of Normal distributions indexed by their means and variances play an important role in the study of real valued random variables, the

multivariate normal distributions are central to the study of random vectors. On $R^d$, a Normal distribution is specified by its probabilty density

$$f(x_1, \ldots, x_d) = k \exp[-\frac{1}{2} Q(x_1 - a_1, \ldots, x_d - a_d)]$$

where $a = (a_1, \ldots, a_d)$ is a location or centering parameter and $Q(x) = Q(x_1, \ldots, x_d)$ is a positive definite quadratic form $Q(x) =< x, Cx >= \sum_{i,j} c_{i,j} x_i x_j$ determined by the symmetric matrix $C = \{c_{i,j}\}$. The normalizing constant $k$ is determined so that

$$\int_{R^d} f(x) dx = 1$$

Clearly by translation and orthogonal rotation the integral can be calculated as

$$k \int_{R^d} \exp[-\frac{1}{2} \sum_{i=1}^{d} \lambda_i y_i^2] dy = (2\pi)^{\frac{d}{2}} \Pi_{i=1}^{d} \frac{1}{\sqrt{\lambda_i}} = 1$$

and

$$k = (\frac{1}{2\pi})^{\frac{d}{2}} \Pi_{i=1}^{d} \sqrt{\lambda_i} = (\frac{1}{2\pi})^{\frac{d}{2}} (\text{Det } C)^{\frac{1}{2}}$$

Here $\lambda_i$ are the eigenvlues of $C$ so that $\Pi_{i=1}^{d} \lambda_i = \text{Det } C$. Since $f(x)$ is symmetric around $x = a$, it is clear that

$$\int_{R^d} x_i f(x) dx = a_i + \int_{R^d} (x_i - a_i) f(x) dx = a_i$$

and thus $\{a_i\}$ are the means of the components $\{x_i\}$. In order to calculate the variances and covariances it is better to calculate the moment generating function

$$\begin{aligned}
&\int_{R^d} \exp[< \theta, x >] f(x) dx \\
&\quad = \int_{R^d} k \exp[-\frac{1}{2} Q(x - a + C^{-1}\theta)] \exp[< a, \theta > + \frac{1}{2} Q(C^{-1}\theta)] \\
&\quad = \exp[< a, \theta > + \frac{1}{2} < C^{-1}\theta, \theta >]
\end{aligned}$$

By differentiatig with respect to $\theta_i$ we can calculate

$$\begin{aligned}
E\big[x_i\big] &= a_i \\
E\big[x_i x_j\big] &= a_i a_j + c_{i,j}^{-1} \\
\text{Cov } x_i x_j &= E\big[x_i x_j\big] - a_i a_j = c_{i,j}^{-1}
\end{aligned}$$

Where $c_{i,j}^{-1}$ is the $i,j$th entry of the inverse $C^{-1}$ of $C = \{c_{i,j}\}$. It is more natural to parametrize the Multivariate Normal Distributions by their means and covariances

$$\{a_i\} = \{E\big[x_i\big]\}$$
$$\Sigma = \{\sigma_{i,j}\} = \{\text{Cov } x_i x_j\}$$

and the density takes the form

$$f(a, \Sigma, x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{\text{ Det } \Sigma}} \exp[-\frac{1}{2} < \Sigma^{-1}(x-a), (x-a) >]$$

We hace assumed that $\Sigma$ is positive definite. In general it only needs to be positive definite. If $\Sigma$ has rank $r < d$ the normal distribution is degenerate and lives on a hyperplane of dimension $r$ and the density can be written down relative to a choice of $r$ coordinates on the hyperplane. If the rank is 0, then $\Sigma = 0$ and the Normal distribution degenrates to a point mass of 1 at the mean.

If $d = 2$, the covariance matrix $\Sigma$ can be written as

$$\begin{bmatrix} \sigma_x^2 & \rho\, \sigma_x \sigma_y \\ \rho\, \sigma_x \sigma_y & \sigma_y^2 \end{bmatrix}$$

with $\Sigma^{-1}$ given by

$$\frac{1}{1-\rho^2} \begin{bmatrix} \frac{1}{\sigma_x^2} & -\frac{\rho}{\sigma_x \sigma_y} \\ -\frac{\rho}{\sigma_x \sigma_y} & \frac{1}{\sigma_y^2} \end{bmatrix}$$

# 20 Testing for Correlation.

If we have $n$ independent observations from a bivariate normal distribution with means $\mu_x, \mu_y$, variances $\sigma_x^2, \sigma_y^2$ and correlation coeffeicient $\rho$ one might want to test that $\rho = 0$ The test naturally will be based on the statistic

$$r = \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{s_x s_y}$$

where $s_x$ and $s_y$ are the sample standard deviations of $x$ and $y$. In order to decide on the critical region we need to determine the distribution of $r$ under

the null hypothesis. Since $r$ is unchanged by any change of origin and/or scale of the observations, we can assume with out loss of generality that $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are two independent sets of independent observations from the standard normal distribution with mean 0 and variance 1. Actually we will assume that $y_1, \ldots, y_n$ are just arbitrary constants and show that the distribution of $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ is $t$ with $n-2$ degrees of freedom no matter what these constants are. Then as long as $x$'s and $y$'s are independent the distribution of $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ will be $t_{n-2}$. If we denote by $a_i = \frac{y_i - \bar{y}}{s_y}$ then $\sum a_i = 0$ and $\sum_i a_i^2 = 1$.

$$r = \frac{\sum a_i x_i}{\sqrt{n} s_x}$$

Let us change coordinates by an orthogonal transformation $z = Sx$ with $S$ given by

$$\begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \ldots & \frac{1}{\sqrt{n}} \\ a_1 & a_2 & \ldots & a_n \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{bmatrix}$$

After the first two rows that form an orthonaormal set of 2 vectors the rest of the matrix is completed to be orthogonal by selectiing the rows to form a complete orthonormal set. In terms of $z_i$, which are again independent standard normals,

$$r = \frac{z_2}{\sqrt{z_2^2 + \cdots + z_n^2}}$$

and

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{z_2\sqrt{n-2}}{\sqrt{z_3^2 + \cdots + z_n^2}}$$

which has a $t$ distribution wth $n-2$ degrees of freedom.

# 21 Large Sample Tests for Correlation.

One can calculate the asymptotic diribution of $r$ for large $n$, in the general case of $\rho \neq 0$. If we define $U_1 = s_x^2, U_2 = s_y^2$ and $U_3 = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}$ and denote by $a_1, a_2$ and $a_3$ their population values $1, 1$ and $\rho$, $\{\sqrt{n}(U_i - a_i)\}$

have a joint normal distribution. The covariance matrix is easily calculated to be

$$A = \begin{bmatrix} 2 & 2\rho^2 & 2\rho \\ 2\rho^2 & 2 & 2\rho \\ 2\rho & 2\rho & 1+\rho^2 \end{bmatrix}$$

From

$$r = \frac{U_3}{\sqrt{U_1 U_2}}$$

we see that $\sqrt{n}(r-\rho)$ is asymptotically normal with variance $< c, Ac >= (1-\rho^2)^2$ with

$$c = (-\frac{\rho}{2}, -\frac{\rho}{2}, 1) = (\frac{\partial r}{\partial U_1}, \frac{\partial r}{\partial U_2}, \frac{\partial r}{\partial U_3})\big|_{(1,1,\rho)}$$

If we consider $z = \frac{1}{2}\log\frac{1-r}{1+r}$ with $z_\rho = \frac{1}{2}\log\frac{1-\rho}{1+\rho}$ then $\sqrt{n}(z-z_\rho)$ is asymptotically normal with mean 0 and variance 1.

# 22 Confidence Intervals.

A confidence interval at level $\alpha$ is a random interval $I$ such that $P_\theta[\theta \in I] \geq \alpha$ for all $\theta$. For example if $X_1, ldots, X_n$ are $n$ independent observations from $N(\mu, \sigma^2)$ an interval of the form $[\bar{x} - ks, \bar{x} + ks]$ contains $\mu$ if $|\frac{\bar{x}-\mu}{s}| \leq k$. The distribution of $\frac{\bar{x}-\mu}{s}\sqrt{n-1}$ is a $t$ with $n-1$ degrees of freedom. We can determine $k_{\alpha,n}$ from th e tables so that $P[|\frac{\bar{x}-\mu}{s}| \leq k] = \alpha$. The interval $[\bar{x} - sk_{\alpha,n}\,, \bar{x} + sk_{\alpha,n}]$ works. Essentially the interval consists of all the possible values of the parameter $\theta$ for which the null hypothesis that the true value is $\theta$ is not rejected at $1-\alpha$ level of significance. In large samples the confidence intervals look like $[\hat{\theta} - \frac{k_\alpha}{\sqrt{n}}\sigma, \theta + \frac{k_\alpha}{\sqrt{n}}\sigma]$, where $k_\alpha$ is determined from the normal table and $\sigma = \sigma(\hat{\theta})$ is the variance of the limiting normal distribution of $\sqrt{n}(\hat{\theta} - \theta)$.