Statistics Revision

Or: What you learned in the pre-requisite course (yes, really!)

Probability

DISCRETE (takes only a finite or countable number of values) random variable X:

$$P(X=x) = p(x) \geq 0$$

p(x) is the probability [mass] function (p.m.f) and

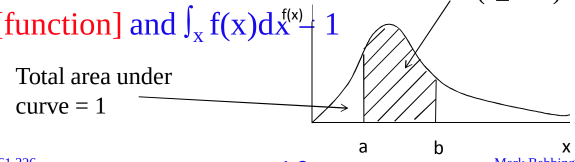
$$\sum_x p(x) = 1$$

Probability

CONTINUOUS (takes only values in an interval) random variable X

$$P(a \leq X < b) = \int_a^b f(x)dx = F(b) - F(a)$$

where $f(x) \geq 0$ is the [probability] density [function] (p.d.f.), $F(x) = P(X \leq x)$ is the distribution [function] and $\int_x f(x)dx = 1$



Bayes Theorem

Suppose X is a random variable (r.v.) taking on values $x_1, x_2, \ldots x_n$, then the Theorem of Total Probability:

$$P(Y=y) = \sum_i P(Y=y|X=x_i)P(X=x_i)$$

where A|B = "A, given that B has occurred".

Bayes Theorem is then

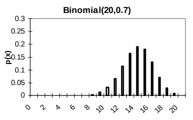
$$P(X=x_i\,|Y=y) = P(Y=y|X=x_i\,)/P(Y=y)$$

which "inverts" the probability.

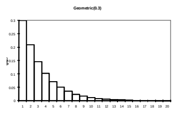
Some Distributions

DISCRETE: independent trials, constant probability of success p

Binomial(n,p): X = number of successes in n trials, $p(x) = (n!/((n-x)!x!))p^x(1-p)^{n-x}$

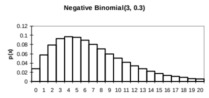


Geometric(p): X = number of failures before first success, $p(x) = p(1-p)^{x-1}$



Negative Binomial(k,p): X = number of failures before kth success,

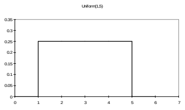
$p(x) = ((k+x-1)!/((k-1)!x!))p^k(1-p)^x$



NB: the letters n,p,k are PARAMETERS, sometimes represented by θ, in which case we write p(x; θ) etc.

More Distributions

CONTINUOUS:

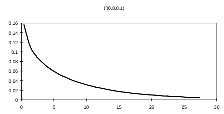


Uniform(a,b): $f(x) = 1/(b-a)$, $a<x<b$

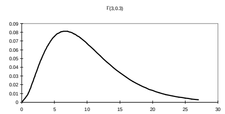
Exponential(λ): $f(x) = \lambda e^{-\lambda x}$, $x > 0$

Gamma(α,λ): $f(x) = \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x} / \Gamma(\alpha)$, $x > 0$

[= "sum" of α Exponential(λ)]



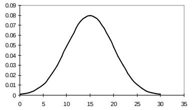
$\alpha < 1$



$\alpha > 1$

Normal (or Gaussian) (μ,σ):

$f(x) = (2\pi\sigma^2)^{-1/2} \exp(-0.5(x-\mu)^2/\sigma^2)$



NB: If $X \sim N(\mu,\sigma)$, then $Z = (x-\mu)/\sigma \sim N(0,1)$

Maximum Likelihood Estimation

Suppose we have n independent observations $x_1, x_2, \ldots x_n$ from a distribution with a p.d.f. $f(x;\theta)$ [or p.m.f. $p(x;\theta)$]. The likelihood is

$$L(\theta; x_1, x_2, \ldots x_n) = \prod_i f(x_i;\theta)$$

The maximum likelihood estimate (MLE) $\hat{\theta}$ is the value of θ maximizing L

Degrees of Freedom

The chi-squared distribution on ν degrees of freedom is the sum of ν squared N(0,1) variables. It measures deviation from the expected.

The t distribution on ν degrees of freedom is very similar to a Normal distribution, just with a larger spread. The difference tends to zero as ν becomes larger.

Basically, degrees of freedom = number of data minus number of constraints. Every estimated parameter is a constraint, and if the sum of the observations has to equal a fixed value, that is another.

Mean and Variance

Mean: $E(X) = \sum_x xp(x) = \int_x x\, f(x)\, dx$

- measure of location

Variance: $V(X) = \sum_x (x - E(X))^2\, p(x)$

$= \int_x (x - E(X))^2 f(x)\, dx$

- measure of spread

Central Limit Theorem (CLT):

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \sim N(E(X), V(X)/n)$$

that is, the mean of a large enough sample has a normal distribution.

Confidence Interval

The method of pivoting: By the CLT

$\bar{X} \sim N(E(X), V(X)/n)$, thus

$$0.95 = P\left(E(X) - 1.96\sqrt{V(X)/n} < \bar{X} < E(X) + 1.96\sqrt{V(X)/n}\right)$$

$$= P\left(\bar{X} - 1.96\sqrt{V(X)/n} < E(X) < \bar{X} + 1.96\sqrt{V(X)/n}\right)$$

gives a 95% confidence interval for E(X). This is an interval that we are 95% confident contains E(X).

Note that this requires knowing V(X), kind of unlikely if we don't know the mean. However, we can replace 1.96 with the corresponding value from the t distribution, and use the sample standard deviation.

For n ~ 60, use 2.0 instead of 1.96.

Hypothesis Testing

Null Hypothesis, $H_N : \theta = \theta_0$

Alternative Hypothesis, $H_A : \theta > \theta_0$ (one sided) or $\theta \neq \theta_0$ (two sided)

Question: How likely (or how "extreme") is the observed data under the null hypothesis. This is the P-value. A small (e.g., < 0.05 at the 5% significance level) P-value corresponds to an unlikely event, which is evidence against the null hypothesis.

Example (Normal Distribution): $H_N : \mu = \mu_0$, $H_A : \mu \neq \mu_0$

Under the null hypothesis, we have a test statistic $Y \sim t_{n-2}$ with observed value $y = (\bar{x} - \mu_0)\sqrt{n}/s$ so the P-value is P(|Y| > y)

(remove the absolute value for a one-sided test).

Bivariate Data

(In DISCRETE notation, CONTINUOUS works similarly)

Joint distribution p(x,y) = P(X=x,Y=y)

$E(XY) = \sum_x \sum_y xyp(x,y)$

Marginal distribution $p(x) = P(X=x) = \sum_y p(x,y)$, => E(X) etc.

Covariance cov(X,Y) = E(XY) – E(X)E(Y)

Correlation $\rho(X,Y) = r = \text{cov}(X,Y)/(V(X)V(Y))^{1/2}$ is a measure of LINEAR dependence. $-1 \leq r \leq 1$

Roughly speaking, for n pairs of data, $r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$

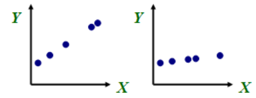
which provides a test for non-zero correlation

With very large sample sizes, weak relationships with low correlation values can be statistically significant!!!

Correlation

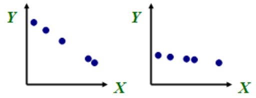
r = 1

A perfect straight line tilting up to the right



r = – 1

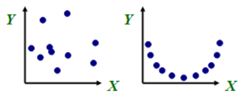
A perfect straight line tilting down to the right



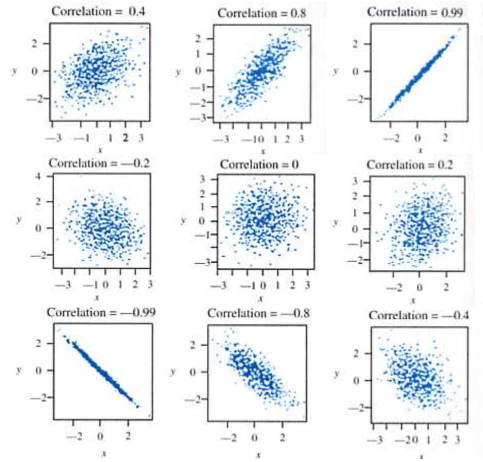
r = 0

No overall tilt

No relationship?



Correlation



Multivariate Data

Joint distribution $p(x_1, x_2,\ldots, x_n) = P(X_1=x_1, X_2=x_2,\ldots, X_n=x_n)$

Covariance matrix $\Sigma = \{\sigma_{ij}\}$ is a symmetric matrix with $\sigma_{ij} = \mathrm{cov}(X_i,X_j)$, i.e., the diagonal is $\sigma_{ii} = V(X_i)$

Multivariate Normal

Let x be the column vector $(x_1\ x_2 \ldots x_n)^T$

Then x has a multivariate normal distribution with mean μ and covariance matrix Σ if

$$f(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

NB: det( ) indicates the determinant, and the argument in the exp( ) is matrix-multiplication.

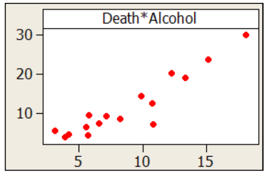
If the covariance matrix Σ is strictly diagonal (all the off-diagonal elements are zero, then the individual components are independent).

Regression

- When the best equation for describing the relationship between variables X and Y is a straight line, the equation is called the *Regression Line*…
- When the relationship between the two variables is linear, a *least squares line* is a useful summary of the relationship, and the *correlation coefficient* is a useful summary of its strength.
- The main purpose of the regression line is to estimate the value of Y at any specified value of X…

Example: (slightly fictitious!)

The graph below shows the *consumption of alcohol* (x) in litres (per year per person aged more than 14 years), and the *death rate* (y) from cirrhosis and alocoholism (per 100 000 population), in sample of 15 randomly selected countries around the world.



1.17

Regression

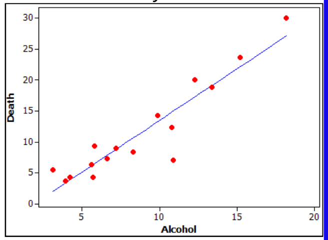
Regression Line for the Sample… $\hat{y} = b_0 + b_1 x$

$\hat{y}$ is spoken as "y-hat," and it is also referred to either as predicted y or estimated y.

b_0 is the *intercept* of the straight line. The intercept is the value of y when x = 0.

b_1 is the *slope* of the straight line. The slope tells us how much of an increase (or decrease) there is for the variable y when the x variable increases by one unit.

The sign of the slope tells us whether y increases or decreases when x increases.



Regression equation:

Death = - 3.216 + 1.666 Alcohol

Slope = 1.666 ⇒ *Death* increases by 1.666 (per 100000), on average, for each increase of 1 unit in *Alcohol*.

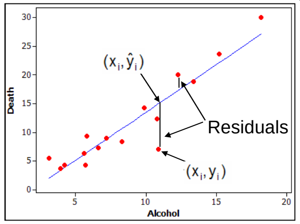
1.18

Regression: Least Squares Line and Formulae

Prediction Errors and Residuals:

- Prediction Error = difference between the observed value of y and the predicted value $\hat{y}$.
- Residual = $(y_i - \hat{y}_i)$
- How good is the fitted line?
- Least Squares Regression Line minimizes the '*sum of squared prediction errors*'…

  SSE = Sum of squared prediction errors or residuals.



$\hat{y}_i = b_0 + b_1 x_i$

Australia has the largest residual…

Formulae for Slope and Intercept:

$$b_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}$$

1.19

Making Inferences…

- In general, Statistical inference examines the question:

  Does the observed characteristic also occur in the population?

  More often, we have no interest in the specific individuals in the data collected. The individuals are 'representative' of a larger population and our main interest is in this underlying population

For a linear relationship…

- What is the slope of the regression line in the population?
- What is the mean value of the *response* variable (y) for individuals with a specific value of the *explanatory* variable (x)?
- What interval of values predict the value of the response variable y for an individual with a specific value of the explanatory variable x?
- Sample vs Population:

  The observed data can be used to determine the regression line for the sample…

  But the regression line for the population can only be imagined…

1.20

Regression Line for the Population…

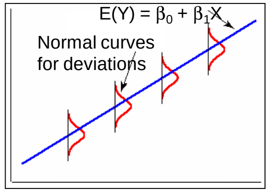
$E(Y) = \beta_0 + \beta_1 X$

- $E(Y)$ represents the *mean* (or *expected value*) of Y for individuals in the population who all have the same X.
- β_0 is the intercept parameter of the straight line in the population…
- β_1 is the slope parameter of the straight line in the population…

  Note that if the population slope β_1 is 0, there is no linear relationship in the population!
- Parameters β_0 and β_1 are estimated using the corresponding (sample) statistics, say, b_0 and b_1.

Assumptions:
"For any x, the distribution of y values is normal…"

$\Rightarrow$ Deviations/residuals from the population regression line have a normal distribution…



Normal Linear Regression for Response…

The most commonly used regression model for the response Y (based on explanatory X) is a "normal linear model".

- Normality - At each value of X, Y has a normal distribution…
- Constant variance - The standard deviation of Y is the same for all values of X…
- Linearity - The mean of Y is linearly related to X…
- Note that, "only the response (Y) is modelled"…

  i.e. a normal linear model tries to explain the variation in Y and does not try to explain the distribution of x-values...

  In *experimental data*, the values of X are fixed by the experimenter, so their distribution is of no interest…

  In *observational data*, the values of X are also usually random and the relationship between X and Y is analysed with a regression model that treats the values of X as constants.

Normal Linear Regression for Response…

Description of the model in terms of a response distribution:

- The normal linear model describes the distribution of Y for any value of X.

  It can be expressed in the form, $Y \sim N(\mu_y, \sigma_y)$

  where $\mu_y = \beta_0 + \beta_1 x$ and $\sigma_y = \sigma$ (for all x, i.e. constant variance)

- An equivalent way to write the same model is,

  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

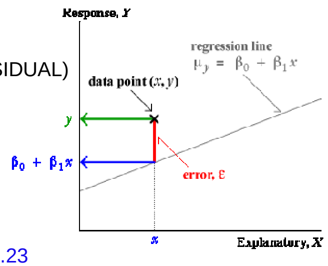
  (i.e. DATA = FIT + RESIDUAL)

  and the residuals (deviations)

  $\varepsilon_i \sim N(0,\sigma^2)$...

  Note that the error, ε_i, for a data point (x_i,y_i) is, $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$

  (i.e. residual or deviation)



Regression Line for the Population…

Least Squares Formulae for Slope and Intercept:

$$b_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}$$

only provide point estimates for β_0 and β_1…

We may wish to compute interval estimates for β_0 and β_1…, say, 95% Confidence Intervals…

$\Rightarrow$ Create a band around the fitted linear regression line that contains about 95% of the values (on the graph) …

Sample-to-sample variability of the least squares estimates means that the least squares slope and intercept in the data are unlikely to be exactly equal to the underlying β_0 and β_1.

$\Rightarrow$ Explore the sampling distribution of b_0 and b_1, the respective sample estimates (statistics) of β_0 and β_1.

Distribution of the Slope and Intercept…

Recall assumptions: "For any x, the distribution of y values is normal…"

i.e. $Y \sim N(\mu_y, \sigma_y)$ and $\varepsilon_i \sim N(0,\sigma^2)$ where $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$…

$\Rightarrow \Rightarrow$

The least squares estimates, b_0 and b_1, have normal distributions that are centered on β_0 and β_1 respectively…

$\Rightarrow b_1 \sim N(\mu_{b1}, \sigma_{b1})$ where $\mu_{b1} = \beta_1$

and

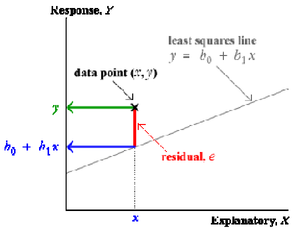
$$\sigma_{b1} = \frac{\sigma}{\sqrt{\sum (x-\bar{x})^2}} = \frac{\sigma}{S_x\sqrt{n-1}}$$

with S_x being the std.dev. of x values

and

$$\hat{\sigma} = \sqrt{\frac{\sum e^2}{n-2}}$$

e^2 are the computed residuals… $\Rightarrow$



Distribution of the Slope and Intercept…

95% Confidence Interval for is… $b_1 \pm t_{n-2}\dfrac{\hat{\sigma}}{S_x\sqrt{n-1}}$

where t_{n-2} is the critical value for a t-distribution with n-2 d.f.

and S_x is the std.dev. of x values…

Example later…

Also, $$\hat{\sigma} = \sqrt{\frac{\sum e^2}{n-2}} = \sqrt{\frac{(n-1)(1-r^2)S_Y^2}{n-2}}$$

where S_Y is std.dev. of y values and r is the correlation coefficient between X and Y…

What affects the accuracy of the least squares slope?

The least squares slope, β_1, has *highest accuracy* when:

the response (or residual) standard deviation, σ, is low

the sample size, n, is large and the spread of x-values is high

Hypothesis Tests …

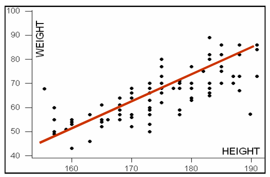
- Importance of zero slope

  If the model's slope is zero, the response distribution does not depend on the explanatory variable...

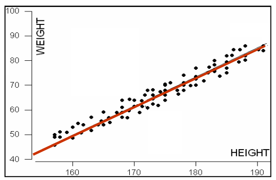
- Strength of relationship vs Evidence for relationship

  It is important to distinguish the strength of a relationship (given by the *correlation coefficient*) and the strength of evidence for existence of a relationship (given by the p-value for the *slope*).

Significant slope, weak correlation



Significant slope, strong correlation



Hypothesis Test: Slope…

Testing for zero slope…

To assess whether the explanatory variable (X) affects the response (Y), we test the hypotheses

H_0: $\beta_1 = 0$ against H_A: $\beta_1 \neq 0$

Test Statistics: (assuming H_0 is true…)

$$t = \frac{b_1 - \beta_1}{s.e.(\beta_1)} = \frac{b_1}{\hat{\sigma}/S_x\sqrt{n-1}} \sim t_{n-2} \text{ distribution}$$

Note: $$\hat{\sigma} = \sqrt{\frac{\sum e^2}{n-2}} = \sqrt{\frac{(n-1)(1-r^2)S_Y^2}{n-2}}$$

Both two-sided and one-sided tests are feasible…

H_0: $\beta_1 = 0$ against H_A: $\beta_1 \neq 0$ or H_A: $\beta_1 > 0$ or H_A: $\beta_1 < 0$

Hypothesis Test: Slope…

Example (Death rate vs Alcohol consumption)…

To assess whether the *Alcohol consumption* (X) affects the *Death rate* (Y) from cirrhosis and alcoholism…

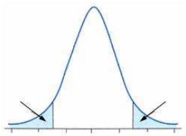
We test the hypotheses H_0: $\beta_1 = 0$ against H_A: $\beta_1 \neq 0$

Compute, $b_1 = 1.666$, $b_0 = -3.216$, $r = 0.933$, $S_x = 4.37$, $S_y = 7.80$, n=16

$$\hat{\sigma} = \sqrt{\frac{\sum e^2}{n-2}} = \sqrt{\frac{(n-1)(1-r^2)S_Y^2}{n-2}} = 2.906$$

Test Statistics: (assuming H_0 is true…)

$$t = \frac{b_1 - \beta_1}{s.e.(\beta_1)} = \frac{b_1}{\hat{\sigma}/S_x\sqrt{n-1}} = 9.70$$



$\Rightarrow$ From t_{14}, p-value <<< 0.001

$\Rightarrow$ We may conclude that there is strong evidence to suggest *Alcohol consumption* significantly affects the *Death rate* from from cirrhosis and alcoholism *in a linear fashion*…

1.29

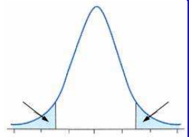
Confidence Interval: Slope…

Example (Death rate vs Alcohol consumption)…

To compute a 95% Confidence Interval for the rate of change in *Death rate with respect to Alcohol consumption* …

95% CI for β_1 is, (with t_{14} from table being 2.145)

$$b_1 \pm t_{n-2}\frac{\hat{\sigma}}{S_x\sqrt{n-1}} = 1.666 \pm 2.145\frac{2.906}{4.37\sqrt{15}}$$



$\Rightarrow$ 1.666 ± 0.3683 = (1.298, 2.034)

& make your own comment here…

Proportion of variation (in Y) explained by the fitted model:

$$= \frac{Var(\hat{Y})}{Var(Y)} = \ldots = r^2$$ and often expressed in %...

In our example, $r^2 = 0.933^2 = 87.1\%$...

$\Rightarrow$ About 87% of variation in death rate can be explained by the fitted *straight line* model indicating a good fit!

1.30

Predicting the Response…

There are two ways in which we may predict/estimate the value of Y for an individual with a particular value of X...

- Suppose we wish to make prediction of the *death rate* (from alcoholism…) of an individual country with an *alcohol consumption rate* of 13.3 (per year per person aged more than 14 years)…
- Alternatively, we may wish to make prediction of the *death rate* of several countries with an (average) *alcohol consumption rate* of 13.3 …
- More generally, we wish to construct a 95% interval of estimates of Y for a particular value of X...

This interval can be interpreted in two equivalent ways:

1. It estimates the central 95% of the values of y for members of population with specified value of x.
2. Probability is .95 that a randomly selected individual from population with a specified value of x falls into the 95% prediction interval.

1.31

Predicting the Response

Recall that the predicted value is, $\hat{y} = b_0 + b_1 x$ for a given x value

Estimating an individual response… (for a particular country with x_0=13.3)

A 95% confidence interval for the 'one' response takes the form

$$\hat{y} \pm t_{n-2}\sqrt{\hat{\sigma}^2 + [s.e.(fit)]^2}$$

where

$$s.e.(fit) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Note the difference…

and t_{n-2} has a t-dist with n-2 d.f…

Estimating mean response (for a number of countries with x_0 = 13.3…)

A 95% confidence interval for the mean response takes the form,

$$\hat{y} \pm t_{n-2}[s.e.(fit)]$$

1.32

Predicting the Response

For the death rate vs alcohol consumption example…

Estimating an individual response… (for a particular country with x_0=13.3)

A 95% confidence interval for the 'one' response takes the form

$$\hat{y} \pm t_{n-2}\sqrt{\hat{\sigma}^2 + \left[\text{s.e.}(\text{fit})\right]^2} \quad = (12.306,\ 25.574)$$

where s.e(fit) = 1.1058 and $\hat{\sigma}$ = 2.906, $t_{14}(0.05)$ = 2.145…

Estimating mean response… (for a number of countries with x_0 = 13.3…)

A 95% confidence interval for the mean response takes the form,

$$\hat{y} \pm t_{n-2}\left[\text{s.e.}(\text{fit})\right] \quad = (16.670,\ 21.210)$$

Note that the CI for individual response ("prediction interval") is wider than the CI for the mean response ("confidence interval")

1.33