Statistics

Week 8: Inference for Proportions (Chapter 9)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Outline

1. Inference for a proportion

2. Chi-squared test
   - Multinomial

Proportions and percentages

Often we are interested in proportions or percentages: the % of people who have a certain opinion or belong to a certain group, the % of successes, the % of defective products, . . .

Question

A survey is conducted to estimate the proportion of people who favour a particular political party. The proportion is to be estimated within a *margin of error* of 3 percentage points, with 95% *confidence*. What sample size should be planned, if the population consists of:

- everyone in this room ($N \approx 30$)?
- everyone in Singapore ($N \approx 5.5$ million)?
- everyone in USA ($N \approx 320$ million)?

Set up

Set up: proportion p of a population has a certain attribute, and we wish to estimate p. We take a random sample, $X_1, X_2, \ldots, X_n$, from the population; they can be treated as iid *Bernoulli* random variables.

An unbiased estimator for p is

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The CLT (or the normal approximation for binomial) tells us that $\hat{p}$ is approximately $N(p,\, p(1-p)/n)$ for *large* n.

Therefore,

$$P\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) \approx 1 - \alpha.$$

When n is large, we can estimate $p(1-p)$ by $\hat{p}(1-\hat{p})$.

Confidence interval

Therefore the $(1-\alpha)$-level, two-sided confidence interval for p is

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Exercise: write down the one-sided CI's.

Remark: it is possible to solve the inequalities on the last slide for p, using the quadratic formula, and obtain a more accurate CI without resorting to the estimate. However, often in practice, the two CI's give very similar results. See textbook Section 9.1.

Sample size

We call $E := z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ the **margin of error**. Then, the minimum sample size for a given margin of error is

$$n = \left(\frac{z_{1-\alpha/2}}{E}\right)^2 \hat{p}(1-\hat{p}).$$

In practice, we may use the conservative value

$$n = \left(\frac{z_{\alpha/2}}{2E}\right)^2,$$

rounded up to the next integer. (Why?)

Exercise: answer the questions on slide 6. Some of the answers may be counter-intuitive, so discuss with your neighbours, and defend your answers if needed.

Hypothesis testing

In inference for a proportion, we usually perform hypothesis testing by computing the p-value.

Unlike in inference for a single sample, this is *not* quite equivalent to constructing a confidence interval.

Example

A basketball player has had a long time average of making 70% of attempted free throws. In the current season he makes 297 out of 396 attempted free throws. Has his free throw average actually improved? Use $\alpha = 0.05$.

Hypothesis testing (continued)

Solution: let $p_0 = 0.7$. We have $H_0 : p = p_0, \;\; H_1 : p > p_0$.

Assume H_0 is true, then $\dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \dfrac{0.75 - 0.7}{\sqrt{0.7 \times 0.3/396}}$ is approximately an observation from a standard normal distribution.

The p-value $= 0.015 < \alpha$, so we reject H_0.

Notice how in computing the p-value, we have used p_0. If we constructed a CI (not the preferred method for hypothesis testing here), then we would have used $\hat{p}$.

Incidentally, the CI is $[0.714, \, 1)$.

Polls – things to watch out for

The media often use polls with small sample sizes, and sensationalize the random fluctuations. Sometimes, the polls are carried out or reported in a way to advance someone's agenda or to sway public opinion.

Internet polling can be inaccurate, especially when the participants are not representative of the population, or if they do not want to disclose their true opinion for whatever reason.

A recent example is the 2016 US presidential election. One poll expert promised to 'eat a bug' if Trump won more than 240 electoral votes (he won 304).

Watch it here

More remarks

There are formulas for computing the power of the test.

There are also ways to perform inferences for comparing two proportions (some of it can also be done via the chi-squared test if the sample size is large).

When the sample size is small, Fisher's exact test can be used.

These are not part of the course. See textbook Sections 9.1 and 9.2.

Outline

1. Inference for a proportion

2. Chi-squared test
   - Multinomial

Multinomial case

We generalize the previous set up: suppose a population is divided into m categories, with proportions $p_1, p_2, \ldots, p_m$ (so $\sum_i p_i = 1$).

We wish to test for $H_0 : p_1 = p'_1,\ p_2 = p'_2, \ldots,\ p_m = p'_m$.

In a sample of size n, let $n_i \in \mathbb{N}$ be the *observed counts* for the ith category, and let $e_i = n\, p'_i$ be the *expected counts*.

Pearson's **chi-squared** statistic is defined as

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - e_i)^2}{e_i},$$

which can be shown, if H_0 is true and $n \to \infty$, to be a χ^2_{m-1} random variable.

Thus, we reject H_0 at significance level α if

$$\chi^2 > \chi^2_{m-1,\,1-\alpha}.$$

Multinomial case, example

In other words, compute

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

and if it is 'too big' (bigger than $\chi^2_{m-1,\,1-\alpha}$), then reject the null.

Example – Mendel's peas

4 characteristic of peas: **S**mooth, **w**rinkled; **Y**ellow, **g**reen.
Offspring of pure **Sw**: $3/4$ **S**, $1/4$ **w**.
Offspring of pure **Yg**: $3/4$ **Y**, $1/4$ **g**.
Their offspring: $9/16$ **SY**, $3/16$ **Sg**, $3/16$ **wY**,$1/16$ **wg**.

Type	**SY**	**Sg**	**wY**	**wg**
Count	315	108	102	31

$m = 4,\ n = 556,\ \chi^2 = 0.604,\ \chi^2_{3,\,0.95} = 7.814.$

Some intuition for the chi-squared test

It is not easy to prove that the χ^2 statistic approximately follows a chi-squared distribution.

Here is a verification for the $m = 2$ case: let X_1 and X_2 denote the random variables for the number of observed in each category. Note that $X_1 + X_2 = n$, and $p_1 + p_2 = 1$.

Then

$$\begin{aligned}\chi^2 &= \frac{(X_1 - p_1 n)^2}{p_1 n} + \frac{(X_2 - p_2 n)^2}{p_2 n} \\ &= \frac{(X_1 - p_1 n)^2}{p_1 n} + \frac{(X_1 - p_1 n)^2}{(1 - p_1)n} \\ &= \frac{(X_1 - p_1 n)^2}{p_1(1 - p_1)n} = \left(\frac{X_1 - p_1 n}{\sqrt{p_1(1 - p_1)n}}\right)^2,\end{aligned}$$

which is the square of one approximately standard normal random variable, i. e. χ^2_1.

Multinomial case, exercise

Exercise – investigate whether the digits of π are random

If they are random, then they would be uniformly distributed. Look at the first 10^{12} digits. Use $\alpha = 0.05$.

Digit	Occurrences
0	99999485134
1	99999945664
2	100000480057
3	99999787805
4	100000357857
5	99999671008
6	99999807503
7	99999818723
8	100000791469
9	99999854780

Statistics

Week 8: Chi-squared Test (Chapter 9)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Outline

1. Chi-squared test
   - Goodness of fit
   - Independence

Goodness of fit – discrete

We can use the chi-squared test to determine whether a specified distribution fits our data!

We first look at an example involving a discrete distribution.

Example – number of passengers per car

Does the data suggest that the numbers of passengers follow a Poisson distribution? Use $\alpha = 0.01$.

Number	0	1	2	3	4
Frequency	678	227	56	28	8

Recall the Poisson distribution has the pmf

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

The (unknown) parameter λ is, conveniently, also the mean.

Discrete example

Refer to the *Excel* file.

- We can *estimate* λ from sample mean, $\hat{\lambda} = 0.456$.
- Use the `exp` and `fact` commands to compute $e_i = n\,p(i)$ (except for e_4).
- **Rule 1**: every time we estimate a parameter, we lose 1 extra degree of freedom.
- **Rule 2**: make sure that each $e_i \geq 5$, in particular, no e_i should be < 1. Try to combine small e_i's with adjacent ones (and do the same to the corresponding observed values).

Final result: $\chi^2 = 72.2$, $\chi^2_{2,\,0.99} = 9.21$.

Continuous example

See *Excel* for an example testing whether some waiting times can be modeled by an exponential distribution (`expon.dist`).

- Since the mean is $1/\lambda$, we can estimate λ by $1/$(sample mean).
- The sample mean calculation can be simplified if we had all the data values.
- There are different ways to group the data, e. g. another way is to make the expected probability constant across the categories.

Final result: $\chi^2 = 6.09$, $\chi^2_{4,\,0.95} = 9.49$.

Two-way tables

The χ^2 statistic can also be computed for two-way tables, to test if the two variables involved are **independent**.
(For more information, see textbook Section 9.4.)

See *Excel* example for Income vs Job Satisfaction, using $\alpha = 0.05$.

- If the null hypothesis (that the variables are independent) is true, then the expected number for each cell in the table is (row sum) $\times$ (column sum) / (grand sum).

- χ^2 is calculated using the same formula, but with a double sum.

- For an $n \times m$ table, the degree of freedom is $(n-1)\,(m-1)$.

Final result: $\chi^2 = 12.0$, $\chi^2_{9,\,0.95} = 16.9$.

Further notes

- You may find the chi-squared test very useful for your project.
- The chi-squared test tends to work well when n is very large (several hundreds, or several thousands).
- A weakness of the chi-squared test is that different groupings of the data may result in different conclusions.

  (The K-S test overcomes this problem.)

Statistics

Week 9: Regression (Chapter 10)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Information

Guest lectures:

Tuesday 21 March, 2–3pm, TT21 (lecture time).

Thursday 23 March, 1–2pm, TT21 (recitation time).

Homework **assignment 2** solutions are available on *eDimension*.

- Q5: it is not convincing to add up the powers of two one-sided tests; refer to the solution.
- Q6 (if using a CI): the CI is *for* σ, *using* s; then checking if σ_0 lies inside it. It is NOT using σ_0, then checking if s lies inside it.
- Q8: there is no way to reduce the problem to testing the variance of *one* population, since we do not know σ_1 or σ_2.

Outline

1. (Simple) linear regression
   - SSE, SST, SSR
2. Multiple (linear) regression

Introduction

Question: how can we construct a line of ‘best fit’ through some data points?

Set up: given n fixed x-coordinates x_i, and n corresponding y-coordinates y_i. A **regression line** is a linear model that describes their relationship.

x is called the predictor/explanatory/independent variable;
y is called the response/ outcome /dependent variable.

We should first make a scatter plot from (x_i, y_i) to check if we have a linear relationship, and if there are outliers.

If a true regression line exists, given by $y = \beta_0 + \beta_1 x$, then we estimate β_0 and β_1 using the **least square** method.

This method is used partly due to mathematical convenience. We do not explore other methods here.

Probabilistic set up

To explain why the data values (x_i, y_i) do not lie perfectly on a straight line, we can think of y_i as the observed value of a random variable Y_i, where

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

and ϵ_i is the random error arising from measurement, variables other than x, etc.

It is common to assume that ϵ_i's are iid *normal* with mean 0 and variance σ^2. This assumption will be useful later when we construct confidence intervals.

An optimization problem

To minimize

$$Q = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i) \right]^2,$$

we compute $\dfrac{\partial Q}{\partial \beta_0}$ and $\dfrac{\partial Q}{\partial \beta_1}$ and set them both to 0.

Denoting the solutions of these equations by $\hat{\beta}_0$ and $\hat{\beta}_1$, we obtain

$$\hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i,$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

These two equations can be routinely solved.

Solution

Notation: let s_x, s_y be the sample standard deviations, let s_{xy} be the sample covariance (`covariance.s` in *Excel*),

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Then we can write the solutions as:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}.$$

The *least square line* is denoted by $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, and is an estimate of the true regression line $y = \beta_0 + \beta_1 x$.

The fitted values are given by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$;
the *residuals* are $e_i := y_i - \hat{y}_i$.

Exercise

In the spreadsheet, find the least square line for the triple jump example using the formulas, and check it against *Excel's* trendline.

Some important terms

The **sum of squared errors (SSE)** is defined to be $\sum_i e_i^2$.

The **sum of squares (total) (SST)** is $\sum_i (y_i - \bar{y})^2 = (n-1)s_y^2$.

It can be shown that

$$\begin{aligned}\textbf{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ &= \qquad \textbf{SSE} \qquad + \qquad \textbf{SSR}.\end{aligned}$$

SSR stands for **sum of squares due to regression**.

Coefficient of determination: $r^2 = \text{SSR}/\text{SST} = 1 - \text{SSE}/\text{SST}$.

An unbiased estimator for σ^2 (of Y_i) is $s^2 = \text{SSE}/(n-2)$, also known as the **mean squared error (MSE)**.

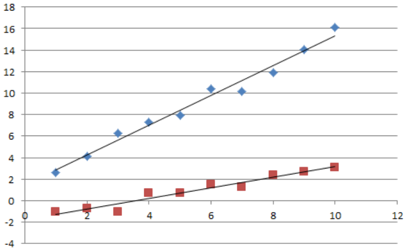
r and r^2

$r^2 \in [0,1]$ can be interpreted as how much of the variation in y can be accounted for by the regression model.

The **correlation coefficient**, $r \in [-1,1]$, is given by

$$r = \frac{s_{xy}}{s_x\, s_y}.$$

Its sign corresponds to the slope of the least square line. $r = \pm 1$ if and only if there is a perfect fit; $r = 0$ means no correlation.



Which least square line has larger *r* (or are they about the same)?

Residuals

Exercise

Compute r^2 for the triple jump example using the formula.

Can the least square line be used to predict the future?

A plot of the *residuals* e_i can be used to check the linearity assumption. For example, a plot which is parabolic in shape indicates the need for an x^2 term.

Rule of thumb: if $|e_i| > 2s$, then the corresponding value may be an outlier.

Example

Investigate the residual plot for the life expectancy data, using `Data Analysis` $\rightarrow$ `Regression`.

Data transformation

If there is a non-linear relationship between x and y, sometimes linear regression can still be used after appropriately transforming the data.

For example, if we suspect $y = \alpha\, x^{\beta}$, then take log of both sides. *Excel* uses `ln` for natural log.

Exercises

(1) What to do if we suspect $y = \alpha\, x^2 + \beta$? $\quad y = \alpha\, e^{\beta x}$?

(2) Interpolate a value in the spreadsheet.

Outline

1. (Simple) linear regression
    - SSE, SST, SSR

2. Multiple (linear) regression

Multiple regression, matrix form

When there are k independent variables, we can construct a least square regression model of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

for the data values $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, $i = 1, 2, \ldots, n$.

Geometrically, this can be a curve, a surface, etc.

Set things up using matrices:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}.$$

Question: what are $\mathbf{X}$ and $\mathbf{y}$ for the life expectancy example?

Solution

We need to minimize

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) = \mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta.$$

This can be done by setting the *gradient* to $\mathbf{0}$, i. e. differentiate the right hand side with respect to each of the β_i's, store the results as a column vector, then set it to the 0 vector.

After manipulation, and using the fact that $\mathbf{X}^T\mathbf{X}$ is symmetric, the result can be simplified to $-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\beta = \mathbf{0}$.

Denoting the solution by $\hat{\beta}$, we obtain

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

A more conceptual proof of this formula can be found in the Math2 Cohort 12 and Cohort 15 slides on projections.

Some properties of multiple regression

- SSE, SST and SSR are defined the same way.
- The formula SSR + SSE = SST still holds.
- $r^2 :=$ SSR/SST.
- r is now the non-negative square root of r^2.
- For polynomial regression, just set the other independent variables as powers of the first one.
- Data transformation works the same way, e. g. for the model $y = \beta_0 \, x_1^{\beta_1} \, x_2^{\beta_2}$, take log of both sides.

Multicollinearity

Beware if some of the independent variables are almost or exactly *linearly dependent*, e. g. income, saving and expenditure. This is sometimes manifested by high correlation between the variables.

If some columns of the matrix $\mathbf{X}$ are linearly dependent, then there exists a non-zero vector $\boldsymbol{v}$ such that $\mathbf{X}\boldsymbol{v} = \mathbf{0}$, so $(\mathbf{X}^T\mathbf{X})\boldsymbol{v} = \mathbf{0}$.

This means $\mathbf{X}^T\mathbf{X}$ is not invertible, making $\hat{\beta}$ impossible to compute. Likewise, if the columns are nearly linearly dependent, then $\mathbf{X}^T\mathbf{X}$ is nearly singular, which causes numerical problems.

Solution: remove a variable that is linearly dependent on the others.

Statistics

Week 10: Regression (Chapter 10 & 11)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Outline

1. Multiple regression
   - Dummy variables
2. Confidence intervals
3. Analysis of variance

Multiple regression

Exercise

In the spreadsheet *regression2 – companies*, use the Data Analysis package to fit a linear model for y in terms of x_1 and x_2.

The regression model can be represented by a plane.

Dummy variables

Sometimes the data contains *categorical* variables, such as gender or seasons. We can encode them using 0's and 1's.

There are different methods of encoding. We demonstrate one method here, using the *Excel* data for triple jump distance vs year and gender.

We set gender $= 0$ for male and 1 for female, and use the model

$$\begin{aligned}\text{distance} &= (\beta_0 + \beta_1 \,\text{gender}) + (\beta_2 + \beta_3 \,\text{gender})\,\text{year} \\ &= \beta_0 + \beta_1 \,\text{gender} + \beta_2 \,\text{year} + \beta_3 \,\text{year} \times \text{gender}.\end{aligned}$$

One advantage of this method is that, when specializing gender to 0 or 1, we recover the least square lines for the male- or female-only data.

Dummy variables, continued

As another example, for the four seasons, we need to introduce *three dummy variables* x_1, x_2, x_3, where:

- $(x_1, x_2, x_3) = (0, 0, 0)$ for spring (chosen as the baseline),
- $(x_1, x_2, x_3) = (1, 0, 0)$ for summer,
- $(x_1, x_2, x_3) = (0, 1, 0)$ for autumn,
- $(x_1, x_2, x_3) = (0, 0, 1)$ for winter.

We do not just use indicator variables here, to avoid multicollinearity.

Again, if ‘interaction’ terms (in *Excel* sheet *sales1*: quarter × season) are included, then specializing the dummy variables gives the individual least square lines. This is a consequence of the underlying matrix algebra.

Outline

1. Multiple regression
    - Dummy variables
2. Confidence intervals
3. Analysis of variance

Set up

In simple linear regression, we can give confidence intervals for β_1 and β_0. Recall the set up

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where ϵ_i are iid normal. We treat the x_i's as fixed, then the Y_i's are normal.

$\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators for β_0 and β_1; in fact they are unbiased.

$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$, so as a random variable, it has distribution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{(n-1)\, s_x^2} = \frac{1}{(n-1)\, s_x^2}\sum_{i=1}^{n}(x_i - \bar{x})\, Y_i,$$

which is a linear combination of normals, and is hence normal.

Calculations

It follows that $\mathsf{E}(\hat{\beta}_1) =$

$$\frac{1}{(n-1)\,s_x^2}\sum_{i=1}^n (x_i-\bar{x})\mathsf{E}(Y_i) = \frac{1}{(n-1)\,s_x^2}\sum_{i=1}^n (x_i-\bar{x})(\beta_0+\beta_1 x_i)$$
$$= \frac{1}{(n-1)\,s_x^2}\sum_{i=1}^n (x_i-\bar{x})\beta_1 x_i = \frac{\beta_1}{(n-1)\,s_x^2}\sum_{i=1}^n (x_i-\bar{x})^2,$$

so $\mathsf{E}(\hat{\beta}_1) = \beta_1$. Likewise,

$$\mathsf{Var}(\hat{\beta}_1) = \frac{1}{(n-1)^2\,s_x^4}\sum_{i=1}^n (x_i-\bar{x})^2\,\mathsf{Var}(Y_i) = \frac{\sigma^2}{(n-1)\,s_x^2}.$$

Similarly tedious computations show that $\hat{\beta}_0$ is normal, with mean β_0 and variance $\frac{\sigma^2}{s_x^2}\Big(\frac{s_x^2}{n}+\frac{\bar{x}^2}{n-1}\Big)$.

Confidence intervals

We estimate σ^2 by $s^2 = \mathsf{SSE}/(n-2)$, which means we will need the t-distribution.

$(1-\alpha)$-level confidence intervals for β_1 and β_0 are, respectively:

$$\hat{\beta}_1 \pm t_{n-2,\,1-\alpha/2}\,\frac{s}{s_x}\,\frac{1}{\sqrt{n-1}},$$

$$\hat{\beta}_0 \pm t_{n-2,\,1-\alpha/2}\,\frac{s}{s_x}\,\sqrt{\frac{s_x^2}{n}+\frac{\bar{x}^2}{n-1}}.$$

These CI's can also be obtained in *Excel*'s Data Analysis → Regression (check the 'confidence level' box). We will check it for the triple jump example.

Note: confidence intervals for β_i in multiple regression can be similarly derived, but involve the diagonal entries of $(\mathbf{X}^T\mathbf{X})^{-1}$ (not in the course; see textbook Section 11.4).

Correlation coefficient

Let ρ denote the true *correlation coefficient* of the random variables X and Y (from which we get the observations (x_i, y_i)). Note that r is just an estimate of ρ. We are interested in testing $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$.

If H_0 is true, then $\rho = 0 = \beta_1$, and one can check that

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{\hat{\beta}_1 - \beta_1}{s/(s_x\sqrt{n-1})},$$

which follows a t-distribution of $(n-2)$ degrees of freedom.

Therefore, we can reject H_0 if

$$\frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} > t_{n-2,\,1-\alpha/2}.$$

Exercise: is $r = 0.5$ always insignificant (with $\alpha = 0.05$)?

Prediction

Suppose we wish to predict the value y^* corresponding to a point x^*. Let $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$.

Then, it can be shown that the $(1-\alpha)$-level two-sided confidence interval for y^* is

$$\hat{y}^* \pm t_{n-2,\,1-\alpha/2}\,\frac{s}{s_x}\sqrt{\frac{s_x^2}{n} + \frac{(x^* - \bar{x})^2}{n-1}}.$$

Outline

1. Multiple regression
   - Dummy variables
2. Confidence intervals
3. Analysis of variance

Predictor variables and r^2

In multiple regression, increasing the number of predictor variables will increase r^2, even if random numbers are used.

This is because each extra predictor variable allows us to decrease the error (in the worst case, just set the new $\hat{\beta}$ to 0 to get the same error, but we are very likely to do better).

As an extreme example, a polynomial regression of degree $(n-1)$ achieves $r^2 = 1$ (where n is the number of data points).

This phenomenon of over-fitting makes r^2 no longer a good measure of how well the model fits the data.

So, how do we pick *useful* predictor variables x_i in our model, and ensure that they have an effect on y?

We first answer a weaker question: how do we know if *any* of the variables affect y?

Analysis of variance (ANOVA)

This question can be answered by ANOVA, the first step of which decomposes the total variability in y into separate components.

We have already done this for multiple (including simple) linear regression: **SST = SSE + SSR**.

Their degrees of freedom are respectively $(n-1)$, $(n-k-1)$, and k, where k is the *number of predictor variables*.

Explanation for the df's: n terms with 1 constraint; n terms with $(k+1)$ parameters estimated; k predictors.

Define **MSE** $=$ SSE$/(n-k-1) = s^2$, and **MSR** $=$ SSR$/k$ (mean squared regression).

Finally, define $F =$ MSR/MSE.

Hypothesis testing using F

(Intuition for SSR having 1 degree of freedom in simple linear regression: observe that $\hat{y}_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x})$.)

In multiple linear regression, it can be shown that under

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0,$$

SSR$/\sigma^2$ and SSE$/\sigma^2$ are both χ^2 random variables.

Therefore, if H_0 is true, then $F =$ MSR/MSE satisfies an $F_{k,\,n-k-1}$ distribution.

If $F > f_{k,\,n-k-1,\,1-\alpha}$, then we can reject H_0, and accept H_1 : at least one of the $\beta_i \neq 0$.

Excel can organize all this information in an ANOVA table.

Statistics

Week 10: Regression (Chapter 10 & 11)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Outline

1. Variable selection

Variable/model selection

If we can reject $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$, then it remains to determine which subset of the predictor variables gives the best model. As mentioned, r^2 is no longer a good measure.

A basic approach is to look at the confidence interval for each β_i, and check if it contains 0 (alternatively, compare the p-value to α).

Example: do this for the *US economy* spreadsheet.

"Essentially, all models are wrong, but some are useful."

– *George Box*

Standardized regression coefficients

Another approach is to compare the effects of each predictor variable on y.

Suppose we have a regression $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. To compare x_1 and x_2 in terms of their effects on y, we cannot compare the sizes $\hat{\beta}_i$ directly, since they may be in different units.

One method is to standardize the data:

$$y_i' = \frac{y_i - \bar{y}}{s_y}, \qquad x_{ij}' = \frac{x_{ij} - \bar{x}_j}{s_{x_j}},$$

then perform the multiple regression.

(In simple linear regression, the new regression line is $\hat{y}' = rx'$.)

Exercise

For the spreadsheet *sales2*, show that x_1 has the larger effect.

Adjusted r^2

For a subset consisting of p of the predictor variables (so $1 \leq p \leq k$), define the adjusted r^2 as

$$r^2_{adj} := 1 - \frac{n-1}{n-1-p}(1-r^2).$$

Then the subset of the x_i's which gives the highest adjusted r^2 can be considered the 'best' model.

This definition is motivated by the observation that a good model should fit the data well using few predictor variables, hence there is a penalty on the number of predictors used.

Exercise

Compute the adjusted r^2 for each of the 3 models for the spreadsheet *sales2*.

More information

The *Akaike information criterion* (AIC) is also commonly used for model selection; it measures the quality of each model relative to the others.

The total number of subsets grows quickly with k, so it is impractical to test for all subsets. *Stepwise regression* (textbook Section 11.7) uses a heuristic for finding a good subset quickly.

AIC and stepwise regression are implemented in R.

Statistics

Week 11: Single Factor Experiments (Chapters 12)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Outline

1. Analysis of single factor experiments

Introduction

Independent samples design allows us to compare two groups. Now we look at techniques for comparing *more than two* groups.

More formally, we look at an experiment which measures a response from more than two groups (or treatments). The treatments are levels of a single treatment factor.

The available experimental units are randomly assigned to each treatment (no matching).

Example: we might want to measure the compression during a crash for small, medium, and large cars.

Set up

Group (or treatment)			
1	2	$\cdots$	k
y_{11}	y_{21}	$\cdots$	y_{k1}
y_{12}	y_{22}	$\cdots$	y_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_{1n_1}	y_{2n_2}	$\cdots$	y_{kn_k}

The group sizes n_i do *not* necessarily equal.

Total sample size: $N = \displaystyle\sum_{i=1}^{k} n_i$.

Sample mean for group i: $\bar{y}_i$.

Sample standard deviation for group i: s_i.

Grand mean: $\bar{\bar{y}} = \dfrac{1}{N} \displaystyle\sum_{i,j} y_{ij}$. Note: this is a double sum.

Set up, continued

We assume that for each group, the response is *normally* distributed, and that all the groups have the *same* variance σ^2 but not necessarily the same mean μ_i.

Exercise

(1) Show that

$$\sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{\bar{y}}\right) = 0.$$

(2) If all the group sizes are the same, give an interpretation of $\bar{\bar{y}}$.

Note: $\sum_i (a_i b_i) \neq \sum_i a_i \ \sum_i b_i$.

Confidence interval

Since each s_i^2 is an estimator of σ^2, we can pool them together to get a better estimate for σ^2:

$$s^2 := \frac{\sum_{i,j}(y_{ij} - \bar{y}_i)^2}{N - k} = \frac{\sum_i (n_i - 1)s_i^2}{\sum_i (n_i - 1)}.$$

Using s, we can write down the $(1-\alpha)$-level confidence interval for μ_i (the true mean of group i):

$$\bar{y}_i - t_{N-k,\,1-\alpha/2}\,\frac{s}{\sqrt{n_i}} \le \mu_i \le \bar{y}_i + t_{N-k,\,1-\alpha/2}\,\frac{s}{\sqrt{n_i}}.$$

The null hypothesis

Our primary interest is in comparing whether the μ_i's are actually different. Set up $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$.

A preliminary test can be carried out using side-by-side box plots.

Can we use confidence intervals to test for H_0? It turns out that even if all the CI's contain a number in common, it is not obvious how strongly this supports H_0.

To test the null hypothesis, it turns out that we can use a tool encountered before: *analysis of variance*.

The idea behind ANOVA

The idea is to compare the variation *between* the groups to the variation *within* each group.

The total variance can (once again) be decomposed into the above two terms.

SST $:= \sum_{i,j}(y_{ij} - \bar{\bar{y}})^2$, df $= N - 1$ (total),

SSE $:= \sum_{i,j}(y_{ij} - \bar{y}_i)^2$, df $= N - k$ (within),

SSA $:= \sum_{i=1}^{k} n_i\,(\bar{y}_i - \bar{\bar{y}})^2$, df $= k - 1$ (between).

The ANOVA identity

SSA is the weighted sum of squared errors between all treatments, and can also be written as $\sum_{i,j}(\bar{y}_i - \bar{\bar{y}})^2$. Its degree of freedom is $(k-1)$ due to the relation in the previous exercise.

A 'large enough' value of SSA would indicate that H_0 is false.

Let **MSA** $=$ SSA$/(k-1)$, **MSE** $=$ SSE$/(N-k) = s^2$, and $F =$ MSA/MSE.

The ANOVA identity

SST = SSA + SSE.

The set up here is just like regression (in fact, it is because the data can be written as a regression model).

Proof of the ANOVA identity

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{\bar{y}})^2 &= \sum_{i,j}(y_{ij}-\bar{y}_i+\bar{y}_i-\bar{\bar{y}})^2\\ &= \sum_{i,j}(\bar{y}_i-\bar{\bar{y}})^2+\sum_{i,j}(y_{ij}-\bar{y}_i)^2+2\sum_{i}(\bar{y}_i-\bar{\bar{y}})\sum_{j}(y_{ij}-\bar{y}_i)\\ &= \sum_{i}n_i(\bar{y}_i-\bar{\bar{y}})^2+\sum_{i,j}(y_{ij}-\bar{y}_i)^2+0.\end{aligned}$$

As in regression, F satisfies a $F_{k-1,\,N-k}$ distribution if H_0 is true.

We can reject H_0 with $(1-\alpha)$ confidence if $F > f_{k-1,\,N-k,\,1-\alpha}$.

Exercises

Use $\alpha = 0.05$ throughout.

(1) Complete all the calculations using the formulas, in the spreadsheet ‘*cars*’.

Check your answers against *Excel*’s `Anova: Single Factor` function.

(2) Complete all the calculations in the spreadsheet ‘*sugar*’; think about how to find SSE and SST.

(3) Construct the ANOVA tables in the spreadsheet ‘*anorexia*’, and answer the question.

Statistics

Week 11: Single Factor Experiments (Chapters 12)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Outline

1. Single factor ANOVA

A summary of ANOVA

Regression

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0.$

If $k = 1$ (only 1 predictor): simple linear regression, easy to test H_0.

If $k > 1$, use ANOVA.

If H_0 is rejected, need to find a good subset of predictors: r^2_{adj}, standardized data, AIC, . . .

Single factor experiment

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k.$

If $k = 2$ (only 2 groups): independent samples design, easy to test H_0.

If $k > 2$, use ANOVA.

If H_0 is rejected, need to find which groups are different: Bonferroni method, Tukey method, . . .

ANOVA table

Obtained in *Excel* under `Data Analysis` → `Anova: Single Factor`.

	SS	**df**	**MS**	F	p-value of F
Between	SSA	$k-1$	MSA	MSA/MSE	1−F.DIST(...)
Within	SSE	$N-k$	MSE		
Total	SST	$N-1$			

This is very similar to ANOVA for regression. In regression, SSA and MSA are replaced by SSR and MSR; the df's are replaced by k, $n-k-1$ and $n-1$.

An informal argument

Why does MSA/MSE follow an F distribution?

We argue informally that regardless to whether H_0 is true,

$$\frac{\text{MSE}}{\sigma^2} = \frac{1}{N-k}\sum_{i,j}\left(\frac{y_{ij}-\bar{y}_i}{\sigma}\right)^2 \sim \frac{\chi^2_{N-k}}{N-k}.$$

When H_0 is true, another informal argument gives

$$\frac{\text{MSA}}{\sigma^2} = \frac{1}{k-1}\sum_{i}\left(\frac{\bar{y}_i-\bar{\bar{y}}}{\sigma/\sqrt{n_i}}\right)^2 \sim \frac{\chi^2_{k-1}}{k-1}.$$

Their ratio follows a $F_{k-1,N-k}$ distribution; their values should be comparable, and a smaller ratio supports H_0 while a larger one supports H_1.

Once H_0 is rejected...

If $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is rejected, we wish to find which treatments have different means. Suppose we naïvely tested each pair: for all $i \neq j$, test a new $H_0 : \mu_i = \mu_j$ by checking if

$$|\bar{y}_i - \bar{y}_j| > t_{N-k,\,1-\alpha/2} \sqrt{\frac{s^2}{n_i} + \frac{s^2}{n_j}},$$

where $s^2 =$ MSE (this is based on independent samples design).

The results would be *incorrect*, because there can be many different pairs, amplifying the probability of observing rare events.

This is another example of the *multiple testing problem*.

Multiple testing problem

If we test across many parameters, then by random chance at least one parameter might show a difference; we are likely to observe coincidences or surprises if we don't specify what to look for in advance.

Although testing for many things at once is fine as an exploratory method, one must use follow-up studies to confirm or refute any patterns that emerge.

Many poorly designed studies fall victim to multiple testing.

Multiple testing problem – examples

- Testing against many distributions for goodness of fit.
- Stonehenge study, published in *Nature* 1963: looked at alignments between 165 features against every rising and setting point for the sun, moon, bright planets and stars.
- Claims of the appearance of constants such as the golden ratio in various settings.
- A Swedish study in 1992 surveyed everyone living within 300m of high-voltage power lines and looked for statistically significant increases in rates of over 800 illnesses.

Bonferroni method

Because there are $m := \binom{k}{2}$ pairs involved in multiple testing , if the type I error for each test is α (and α is small enough), then the overall error is about $m\alpha$.

To fix this, we insist that the error for *each* test has to be α/m, so that the overall error is about α.

This approach is known as the *Bonferroni method*: we reject $H_0 : \mu_i = \mu_j$ if

$$|\bar{y}_i - \bar{y}_j| > t_{N-k,\, 1-\alpha/(2m)} \sqrt{\frac{s^2}{n_i} + \frac{s^2}{n_j}}.$$

There are other methods to check whether two treatments are different (textbook, Section 12.2); no method is perfect. Bonferroni works well for small k, but can be too conservative.

ANOVA – further information

ANOVA is very widely used, but also widely abused.

ANOVA is reasonably robust against violations of its assumptions, namely,

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

where ϵ_{ij} are iid $N(0, \sigma^2)$ random variables.

The *residuals* $e_{ij} := y_{ij} - \bar{y}_i$ can be used to test for normality, via a Q-Q plot.

There are tests to check if the variances are equal, and methods to transform the data if they are not (Section 12.1.3).

Exercise

Check the normality assumption for the '*anorexia*' example.

Use the Bonferroni method to determine which treatment is better.

Statistics

Week 12: Two-Factor Experiments (Chapters 13)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF
TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Randomized block design

ANOVA can also be used in randomized block design (not in this course, see textbook Section 12.4).

For example, to compare the flight distances of 3 types of golf balls, we can get each golfer to hit one ball of each type in random order.

The golfer is the blocking (noise) factor. F values can then be used to test for treatment effects and blocking effects.

Outline

1. Two-factor experiments

Introduction

We now consider experiments with two factors, which we call A and B. Factor A has $a \geq 2$ levels, and factor B has $b \geq 2$ levels.

Example: measure the compression during a crash, vs car size (factor A) and car speed (factor B).

We consider all treatment combinations (in total, ab of them). For each treatment combination, we make n observations (called replicates).

See *Excel* sheet *anova2* for an example.

We omit the mathematical details in this course (but see textbook Section 13.1 if you are interested).

Layout

	Factor B levels			
Factor A levels	1	2	$\cdots$	b
1	y_{111} $\vdots$ y_{11n}	y_{121} $\vdots$ y_{12n}	$\cdots$	y_{1b1} $\vdots$ y_{1bn}
2	y_{211} $\vdots$ y_{21n}	y_{221} $\vdots$ y_{22n}	$\cdots$	y_{2b1} $\vdots$ y_{2bn}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a11} $\vdots$ y_{a1n}	y_{a21} $\vdots$ y_{a2n}	$\cdots$	y_{ab1} $\vdots$ y_{abn}

Interactions

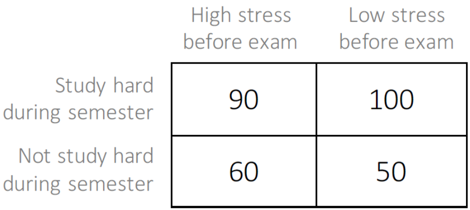
We wish to test whether the treatments have the same effect. Do we really need a new method to analyze this experimental set-up, or can we just test for each factor separately?

It turns out that we cannot just test each factor separately, because there might exist **interactions**: when the effects of one factor depend on the levels of the other factor.

To visualize interactions, we can create a *line chart* for each level of Factor A. In each line chart, plot the cell means against the levels of Factor B.

Interactions – example

(From Recitation 1) suppose we have the following cell means:



	High stress before exam	Low stress before exam
Study hard during semester	90	100
Not study hard during semester	60	50

Exercise

If the lines are almost parallel, then what does that say about the level of interaction?

ANOVA

There are three natural null hypotheses to look at:

H_{0A} : the levels of Factor A all have the same mean

H_{0B} : the levels of Factor B all have the same mean

H_{0AB} : there is no interaction between Factors A and B

These null hypotheses are tested by looking at the values of F_A, F_B, and F_{AB} (computable in *Excel*) respectively.

If each null hypothesis is true, then it can be shown that the corresponding F follows an F distribution.

Excel

One should look at F_{AB} first. If H_{0AB} is rejected, then the presence of interaction means that the other two hypothesis tests are no longer very meaningful.

In *Excel*, go to `Data Analysis` → `Anova: Two-Factor With Replication`.

Include the column and row headings in the input range, and enter the correct 'Rows per sample'.

Exercise

Do the analysis for *anova2*.

Statistics

Week 12: Other Statistical Methods (Chapter 14)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF
TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Information

ANOVA and HW3 solutions will be available on eDimension.

Homework assignment 4 will be available soon.

Please complete the **course survey**!

We now study *non-parametric* tests, which are tests that do not assume that the samples come from a nice distribution (e. g. normal). Bootstrapping and permutation tests are examples of non-parametric tests.

Outline

1. Sign test
2. Signed rank test
3. Runs test

Test for the median

The **sign test** is a simple test for the median. As the median is robust, it can be a better measure of centre than the mean (and is often used for income, property prices, etc).

We would like to test $H_0 : m = m_0$, where m is the true but unknown median and m_0 is a specific value.

Given data points $x_1, x_2, x_3, \ldots$, we first ignore any x_i that equals m_0. Let n be the number of data points that differ from m_0.

Next, we *count the number of* x_i's that exceed m_0, and call that number s_+. (We may also use $s_- := n - s_+$.)

Idea: reject H_0 if s_+ is 'too different' from $n/2$.

Sign test

If H_0 were true, then each $x_i > m_0$ with probability $1/2$. So S_+ follows a *binomial* distribution.

For the one-sided alternative $H_1 : m > m_0$, the p-value of s_+ is

$$\sum_{i=s_+}^{n} \binom{n}{i} 2^{-n}.$$

Reject H_0 if the p-value $< \alpha$.

The two-sided test is similar.

This is a *non-parametric* test, since it makes no assumption about the underlying distribution.

Example – hypothesis test

For the data values

$$6,\ 8,\ 9,\ 5,\ -7,\ 5,\ 3,\ -3,\ 0,\ -12,\ 3,\ 1,$$

test $H_0 : m = 0$ vs $H_1 : m > 0$, using $\alpha = 0.1$.

Answer: $s_+ = 8$, so the p-value is $\displaystyle\sum_{i=8}^{11} \binom{11}{i} 2^{-11} = 0.113$; do not reject H_0.

In *Excel*: use `binom.dist`.

Confidence interval

We can provide a crude confidence interval for the median m.

Suppose we want to find a 95% confidence interval (which turns out to be not possible, but we can get close).

We first work with S_+. Make each tail probability as close to 2.5% as possible:

$$\sum_{i=0}^{2} \binom{12}{i} 2^{-12} = \sum_{i=10}^{12} \binom{12}{i} 2^{-12} \approx 0.0193.$$

So with 96% probability, $3 \leq S_+ \leq 9$.

Next, we convert this into a 96% confidence interval for m, which is $[-3, 6)$.

Outline

1. Sign test
2. Signed rank test
3. Runs test

Wilcoxon signed rank test

The Wilcoxon **signed rank test** takes into account the values of $d_i := x_i - m_0$, which makes it more *powerful*. However, it requires the extra assumption that the population distribution is *symmetric*.

Procedure: (1) *Rank* the d_i's in terms of absolute values (the smallest receives rank 1, etc). In case of ties, use the *average* rank.

(2) Let w_+ be the sum of the ranks of the positive d_i's.

(3) For moderate sized n, under $H_0 : m = m_0$, W_+ is approximately normal with

$$\mu = \frac{n(n+1)}{4}, \quad \sigma^2 = \frac{n(n+1)(2n+1)}{24}.$$

(n is the number of data values that differ from m_0.) Reject H_0 if w_+ is too many σ away from μ.

Proof

If H_0 were true, then the d_i which receives the kth rank is positive with probability $1/2$. Therefore

$$W_+ = \sum_{k=1}^{n} k\, X_k,$$

where X_k are iid Bernoulli random variables with parameter $1/2$. From this we can calculate its expectation and variance. Normality follows from a stronger form of the CLT.

Note: there is a test that uses a similar idea, called the Wilcoxon rank sum test (textbook Section 14.2, not in the course), which checks if two independent samples come from the same population.

Example

See spreadsheet *rank* for an example. Use *Excel*'s `rank.avg` and `if` functions.

In this example, $\mu = 60$, $\sigma^2 = 310$, $w_+ = 101$. The p-value is

$$P(Z > \big(100.5 - 60)/\sqrt{310}\big) \approx 0.0107.$$

We have used the *continuity correction*.

Note that the *sign test* would not have rejected H_0.

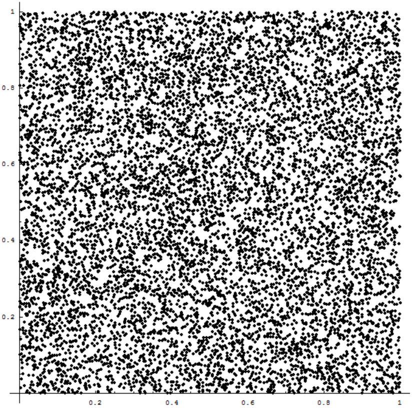
Outline

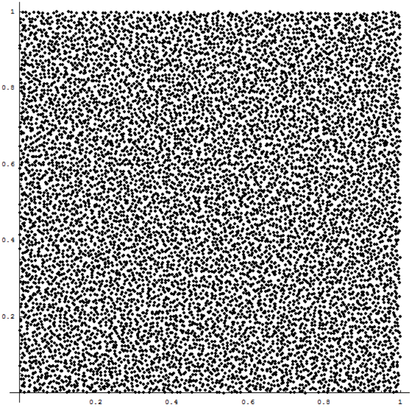
1. Sign test

2. Signed rank test

3. Runs test

Which one of the following pictures is random?





Number of runs

Suppose we observe a sequence of events, consisting of n H's and m T's (modeled as coin tosses). P(H) is *unknown*.

We would like to know if the elements of the sequence are *independent*. This could be used to check if a sequence of coin tosses were made up, or if it were random.

To do so, we count the number of **runs**.

H, T, H, H, H, H, H, T, T, H, T, T, T, H, T,
H, H, T, T, H, T, T, T, H, H, H, H, T, T, H.

T, H, H, T, H, T, H, T, H, H, T, H, H, T, H,
T, T, H, H, T, H, T, T, H, T, H, H, T, H, T.

In the first sequence, there are 8 runs of H's, in the second sequence there are 11's runs of H. (Runs of T's are *within 1* of runs of H's.)

The Wald-Wolfowitz runs test

Under H_0 : *H's and T's are independently drawn from the same distribution*, we have:

The *total* number of runs is approximately normal with

$$\mu = \frac{2mn}{m+n} + 1, \qquad \sigma^2 = \frac{(\mu - 1)(\mu - 2)}{m + n - 1},$$

where n is the number of H's and m is the number of T's .

Thus, if the total number of runs is too many σ away from μ, then we reject H_0.

Exercise

For the second sequence of coin tosses, compute the p-value for the total runs, and determine if the sequence was made up. Use $\alpha = 0.05$, and do not forget the continuity correction.

Proof (probability for runs of H's)

We give some reasons as to why the runs are normally distributed. As a simplification, the proof here *only* considers the runs of H's.

If we assume that H's and T's are independently drawn from the same distribution, then any of the $\binom{m+n}{n}$ arrangements are equally likely. We want to know the probability of getting r runs of H's.

Consider the H's first, there are $\binom{n-1}{r-1}$ ways to break them up into r runs.

To distribute the r runs among the m T's, there are $\binom{m+1}{r}$ ways.

The runs of H's thus satisfy a *hypergeometric* distribution:

$$P(r \text{ runs of H's}) = \frac{\binom{n-1}{n-r}\binom{m+1}{r}}{\binom{m+n}{n}}.$$

Proof (normal approximation)

For moderate sized m and n, this can be approximated using a *normal* distribution with the same mean and variance:

$$\mu = \frac{n(m+1)}{m+n}, \qquad \sigma^2 = \frac{n(n-1)m(m+1)}{(m+n-1)(m+n)^2}.$$

(So for $m \approx n$, we expect around $n/2$ runs of H; anything too different allows us to reject H_0.)

Note that the Wald-Wolfowitz runs test also takes into account the runs of T's, but the proof idea is similar.

Applications

The runs test is very useful, as it makes no assumptions about P(H), and there are many ways to convert data into a sequence of H's and T's:

- Wins and losses,
- Whether data values are above/below the median,
- For integer data, whether the values are even/odd,
- Whether data values over/under-fit a distribution.

There are also tests based on the longest runs (e. g. in N tosses of a fair coin, the longest run of H's or T's is very likely to be around $\log_2(N) - 0.5$).

Statistics

Week 13: Maximum Likelihood (Chapter 15)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Information

- **Assignment 4** out today, due next Monday.
- Updated slides as a single pdf will be available on *eDimension*.
- Tomorrow: revision.
- Anyone doing Project Option 2?
- Please complete the **course survey**.
- **Exam**: 9–11am, Friday 28 April.

  – Only the second half will be explicitly tested, though you need to know the concepts learned in the first half.

  – An A4 sheet with handwritten notes on both sides, and a non-programmable calculator are allowed.

  – You need to understand how least square regression works.

Outline

1. Maximum likelihood

Parameter estimation, revisited

We talked about estimators, in particular unbiased estimators, but the ways in which we constructed them have been ad hoc.

For example, in Week 8 we had to estimate the parameter λ in a Poisson distribution. We used the fact that λ is also the expectation, and then estimated it using the sample mean.

What if we didn't know how to relate λ to the expectation? We now describe a general approach to parameter estimation.

The idea is to pick the value of the parameter which *maximizes the probability* of observing our data.

For example, suppose we want to estimate the parameter p of a Bernoulli distribution. A random sample of size 5 drawn from this distribution reads: 1, 1, 1, 1, 1.

Which of these is most likely? $p = 0;\ p = 1/2;\ p = 2/3;\ p = 1.$

Likelihood function

Given a probability distribution, let θ be a parameter to be estimated, and denote the probability density (or mass) function by $f(x|\theta)$.

The joint density for n iid random observations, $x_1, x_2, \ldots, x_n$, is

$$L(\theta) := \prod_{i=1}^{n} f(x_i|\theta).$$

$L(\theta)$ is called the **likelihood function** of θ.

If the maximum of this function occurs at $\hat{\theta}$, then $\hat{\theta}$ is the value of the parameter which maximizes the probability of observing our data, and we will use it as our estimate for θ.

Maximum likelihood estimate

$\hat{\theta}$ is called the **maximum likelihood estimate**. In many cases, we can find $\hat{\theta}$ by solving for

$$\frac{\mathrm{d}}{\mathrm{d}\theta} L(\theta) = 0, \quad \text{or equivalently,} \quad \frac{\mathrm{d}}{\mathrm{d}\theta} \log(L(\theta)) = 0.$$

('log' stands for natural log.)

Example – Poisson

$$L(\theta) = e^{-n\theta}\, \theta^{x_1+x_2+\cdots+x_n}\, \frac{1}{x_1!x_2!\cdots x_n!},$$

solving for $\frac{\mathrm{d}}{\mathrm{d}\theta} \log(L(\theta)) = 0$, we find that $\hat{\theta} = \bar{x}$.

We can use the *2nd derivative* to check that it is a maximum.

Maximum likelihood estimate – applications

- Maximum likelihood is used to find the coefficients in *logistic regression*.

- AIC $= 2k - 2\log(\hat{L})$, where $k =$ the number of estimated parameters in the model, and $\hat{L}$ is the maximized value of the likelihood function.

- Maximum likelihood is sometimes used to introduce *Bayesian statistics*, which is an alternative approach to the *frequentist statistics* taught in this course.

Exercises

1. (Exponential) If $f(x|\theta) = \theta e^{-\theta x}$, estimate θ based on x_i.

2. (Normal) Find the maximum likelihood estimates for μ and σ^2 in a normal distribution, whose pdf is given by

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$

Are the estimators unbiased?

3. (Uniform) Suppose $x_1 = 1.2, x_2 = 3.5, x_3 = 2.7$ is a random sample from a *continuous uniform distribution* over $[0,\ \theta]$.

Then $f(x|\theta) = 1/\theta$ if $0 \leq x \leq \theta$, and 0 otherwise. Find $\hat{\theta}$.

Generalize your result. Does this give an unbiased estimator?