Lecture Aims

- Compare observed counts to a hypothesized distribution
- Test for association in two-way tables

What is ‘normal’?

What is 'normal'?

'Normal' is a technical adjective used in various areas of mathematics to mean different things.

What is 'normal'?

'Normal' is a technical adjective used in various areas of mathematics to mean different things.

The Normal distribution is only one of many for describing statistical models, though it certainly is important because of the Central Limit Theorem.

What is ‘normal’?

‘Normal’ is a technical adjective used in various areas of mathematics to mean different things.

The Normal distribution is only one of many for describing statistical models, though it certainly is important because of the Central Limit Theorem.

A *normal vector* is a vector that is perpendicular to another vector or surface.

Normal numbers

A real number is *normal* if its infinite sequence of digits is distributed uniformly.

Normal numbers

A real number is *normal* if its infinite sequence of digits is distributed uniformly.

Is $\sqrt{2}$ a normal number?

$\sqrt{2}$

The first 1000 digits of $\sqrt{2}$ are

1.41421356237309504880168872420969807856967187537694
80731766797379907324784621070388503875343276415727
35013846230912297024924836055850737212644121497099
93583141322266592750559275579995050115278206057147
01095599716059702745345968620147285174186408891986
09552329230484308714321450839762603627995251407989
68725339654633180882964062061525835239505474575028
77599617298355752203375318570113543746034084988471
60386899970699004815030544027790316454247823068492
93691862158057846311159666871301301561856898723723
52885092648612494977154218334204285686060146824720
77143585487415565706967765372022648544701585880162
07584749226572260020855844665214583988939443709265
91800311388246468157082630100594858704003186480342
19489727829064104507263688131373985525611732204024
50912277002269411275736272804957381089675040183698
68368450725799364729060762996941380475654823728997
18032680247442062926912485905218100445984215059112
02494413417285314781058036033710773091828693147101

Observed Counts

Digit	0	1	2	3	4	5	6	7	8	9
Observed	108	98	109	82	100	104	90	104	113	92

Clicker Question

Channel 41

We want to test the null hypothesis that $\sqrt{2}$ is normal. If so, how many counts would we expect for each digit?

1. 90
2. 100
3. $\frac{1000}{\sqrt{2}} = 707.1$

Clicker Question

Channel 41

Given 1000 digits, did we need to look at the observations to calculate the expected counts in the previous question?

1. Yes
2. No

Expected Counts

Digit	0	1	2	3	4	5	6	7	8	9
Observed	108	98	109	82	100	104	90	104	113	92
Expected	100	100	100	100	100	100	100	100	100	100

Expected Counts

Digit	0	1	2	3	4	5	6	7	8	9
Observed	108	98	109	82	100	104	90	104	113	92
Expected	100	100	100	100	100	100	100	100	100	100

How should we measure the differences between the observed and expected counts?

χ^2 Statistic

As usual we compare things with sums of squared deviations but here we make them relative to the size of the expected count.

χ^2 Statistic

As usual we compare things with sums of squared deviations but here we make them relative to the size of the expected count.

This gives the χ^2 statistic

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

χ^2 Statistic

As usual we compare things with sums of squared deviations but here we make them relative to the size of the expected count.

This gives the χ^2 statistic

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

Here we have

$$\chi^2 = \frac{(108 - 100)^2}{100} + \frac{(98 - 100)^2}{100} + \cdots + \frac{(92 - 100)^2}{100} = 8.38$$

Clicker Question

Channel 41

If the null hypothesis is true then this χ^2 statistic has a χ^2 *distribution*. What are the degrees of freedom of this distribution?

1. 1
2. 9
3. 99
4. 999

χ^2 Statistic

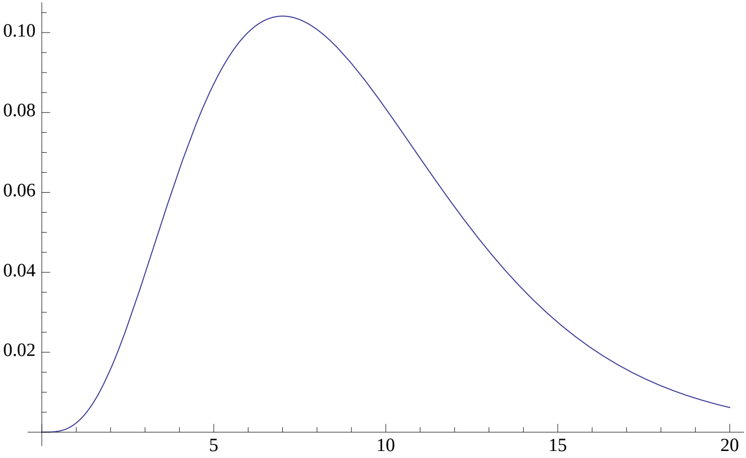
Is a χ^2 value of 8.38 significant evidence that the digits of $\sqrt{2}$ are not uniformly random?

χ^2 Statistic

Is a χ^2 value of 8.38 significant evidence that the digits of $\sqrt{2}$ are not uniformly random?

If the null hypothesis is true then this statistic has a χ^2 distribution with $k-1$ degrees of freedom where k is the number of categories.

χ^2_9 Density Function



χ^2 Statistic

Our p-value here is $P[\chi^2_9 \geq 8.38] = 0.496$.

χ^2 Statistic

Our p-value here is $P[\chi^2_9 \geq 8.38] = 0.496$.

Alternatively we can give a range from the χ^2 tables.

$$P[\chi^2_9 \geq 8.38] > 0.25$$

χ^2 Statistic

Our p-value here is $P[\chi^2_9 \geq 8.38] = 0.496$.

Alternatively we can give a range from the χ^2 tables.

$$P[\chi^2_9 \geq 8.38] > 0.25$$

Either way there is no evidence to suggest $\sqrt{2}$ is not normal.

22.4 Two-way Tables

Recall the ear infection data:

Syrup	Infection	No Infection	Total
Placebo	68	97	165
Xylitol	46	113	159
Total	114	210	324

22.4 Two-way Tables

Recall the ear infection data:

Syrup	Infection	No Infection	Total
Placebo	68	97	165
Xylitol	46	113	159
Total	114	210	324

If there was no association between xylitol and ear infection, what counts would we expect to see?

Expected Counts

Ignoring the groups, the proportion who had an ear infection was $\frac{114}{324}$.

Expected Counts

Ignoring the groups, the proportion who had an ear infection was $\frac{114}{324}$.

Ignoring the infections, the proportion in the xylitol group was $\frac{159}{324}$.

Expected Counts

Ignoring the groups, the proportion who had an ear infection was $\frac{114}{324}$.

Ignoring the infections, the proportion in the xylitol group was $\frac{159}{324}$.

If the null hypothesis is true then these outcomes should be independent and so we could multiply the proportions together to estimate the count we would expect in the corresponding cell:

$$\frac{114}{324} \times \frac{159}{324}$$

Expected Counts

Ignoring the groups, the proportion who had an ear infection was $\frac{114}{324}$.

Ignoring the infections, the proportion in the xylitol group was $\frac{159}{324}$.

If the null hypothesis is true then these outcomes should be independent and so we could multiply the proportions together to estimate the count we would expect in the corresponding cell:

$$\frac{114}{324} \times \frac{159}{324} \times 324 = 55.9.$$

Expected Counts

Ignoring the groups, the proportion who had an ear infection was $\frac{114}{324}$.

Ignoring the infections, the proportion in the xylitol group was $\frac{159}{324}$.

If the null hypothesis is true then these outcomes should be independent and so we could multiply the proportions together to estimate the count we would expect in the corresponding cell:

$$\frac{114}{324} \times \frac{159}{324} \times 324 = 55.9.$$

We can do this for all cells in the table.

Expected Counts

Repeating this for all cells gives the expected counts under H_0 : "independence between Syrup and Infection".

Syrup	Infection	No Infection	total
Placebo	$\frac{165\times114}{324} = 58$	$\frac{165\times210}{324} = 107$	165
Xylitol	$\frac{159\times114}{324} = 56$	$\frac{159\times210}{324} = 103$	159
Total	114	210	324

χ^2 Statistic

We measure the difference between the observed and expected values using a χ^2 statistic, with

$$x^2 = \frac{(68-58)^2}{58} + \cdots + \frac{(113-103)^2}{103} = 5.41$$

χ^2 Statistic

We measure the difference between the observed and expected values using a χ^2 statistic, with

$$x^2 = \frac{(68 - 58)^2}{58} + \cdots + \frac{(113 - 103)^2}{103} = 5.41$$

If there is no association then this statistic has a χ^2 distribution with degrees of freedom

$$(\# \text{ rows} - 1)(\# \text{ columns} - 1).$$

Ear infections

The p-value is thus

$$0.01 < P[\chi^2_1 \geq 5.41] < 0.025,$$

evidence that the variables are not independent and so that there is an association between xylitol and ear infection.

Assumptions for χ^2 test

The χ^2 distribution is an approximation to the discrete distribution.

Assumptions for χ^2 test

The χ^2 distribution is an approximation to the discrete distribution.

For good approximation require all expected values to be at least 1 and 80% of them to be at least 5.

Assumptions for χ^2 test

The χ^2 distribution is an approximation to the discrete distribution.

For good approximation require all expected values to be at least 1 and 80% of them to be at least 5.

For smaller samples use *Fisher's exact test*.

Five steps: χ^2 for independence

1. Define the null and the alternative hypotheses:
   H_0 : "independence between treatment and ear infection" and H_1 : "association"
2. Statistic test:
   $$X^2 = \frac{\sum(OBSERVED - expected)^2}{expected} \sim \chi^2((r-1)\times(c-1))$$

Five steps: χ^2 for independence

1. Define the null and the alternative hypotheses:
   H_0 : "independence between treatment and ear infection" and H_1 : "association"
2. Statistic test:
   $$X^2 = \frac{\sum(OBSERVED - expected)^2}{expected} \sim \chi^2((r-1)\times(c-1))$$
3. Realisation of Statistic test:
   $$x^2_{obs} = \frac{\sum(observed - expected)^2}{expected}, \quad \text{assuming } H_0 \text{ true}$$

   where

Syrup	Infection	No Infection	total
Placebo	$\frac{A\times C}{E}$	$\frac{A\times D}{E}$	A
Xylitol	$\frac{B\times C}{E}$	$\frac{B\times D}{E}$	B
Total	C	D	E

Five steps: χ^2 for independence

(4) Computation of the p-value:

$$\text{p-value} = P(X^2 \geq x^2_{obs}) \text{ where } X^2 \sim \chi^2((r-1)\times(c-1))$$

Five steps: χ^2 for independence

(4) Computation of the p-value:

$$\text{p-value} = P(X^2 \geq x^2_{obs}) \text{ where } X^2 \sim \chi^2((r-1) \times (c-1))$$

(5) Conclude:
p-value $\leq \alpha$ (=0.05), evidence against H_0, we reject H_0.
p-value $> \alpha$, no evidence against H_0.

Data Reduction

Recall that the p-value is a transformation of the data:

Data Reduction

Recall that the p-value is a transformation of the data:

Data

Data Reduction

Recall that the p-value is a transformation of the data:

$$\text{Data} \longrightarrow \overline{X}$$

Data Reduction

Recall that the p-value is a transformation of the data:

$$\text{Data} \longrightarrow \overline{X} \longrightarrow T$$

Data Reduction

Recall that the p-value is a transformation of the data:

$$\text{Data} \longrightarrow \overline{X} \longrightarrow T \longrightarrow P$$

Data Reduction

More generally we have seen

$$\text{Data} \quad \longrightarrow \quad Z, T, F, \chi^2 \quad \longrightarrow \quad P$$

Clicker Question

Channel 41

What is the method to use in this context ?

1. T-test
2. Z-test
3. linear model
4. logistic model
5. Anova method
6. χ^2 test
7. other

Summary: one population

1. Continuous variable: weight, height, pulse rate

Summary: one population

1. Continuous variable: weight, height, pulse rate
   test or confidence interval for the population mean μ

Summary: one population

1. Continuous variable: weight, height, pulse rate
   test or confidence interval for the population mean μ
   by T-test, Z-test (Normal distribution)

Summary: one population

1. Continuous variable: weight, height, pulse rate
   test or confidence interval for the population mean μ
   by T-test, Z-test (Normal distribution)
2. Binary variable: Disease status (HIV), gender (female, male)

Summary: one population

1. Continuous variable: weight, height, pulse rate
   test or confidence interval for the population mean μ
   by T-test, Z-test (Normal distribution)
2. Binary variable: Disease status (HIV), gender (female, male)
   test or confidence interval for the population proportion p

Summary: one population

1. Continuous variable: weight, height, pulse rate
   test or confidence interval for the population mean μ
   by T-test, Z-test (Normal distribution)
2. Binary variable: Disease status (HIV), gender (female, male)
   test or confidence interval for the population proportion p
   by Z test (Normal distribution)

Summary: Continuous $\times$ binary

Example: Y: height, pulse rate and X: gender, Treatment (only two: placebo and new treatment)

Summary: Continuous $\times$ binary

Example: Y: height, pulse rate and X: gender, Treatment (only two: placebo and new treatment)
$\hookrightarrow$ two populations defined by the binary variable.

Summary: Continuous × binary

Example: Y: height, pulse rate and X: gender, Treatment (only two: placebo and new treatment)
$\hookrightarrow$ two populations defined by the binary variable.

1. Comparison of two means:
   test or confidence interval for the population mean μ

Summary: Continuous $\times$ binary

Example: Y: height, pulse rate and X: gender, Treatment (only two: placebo and new treatment)
$\hookrightarrow$ two populations defined by the binary variable.

1. Comparison of two means:
   test or confidence interval for the population mean μ
   by T-test, Z-test (Normal distribution)

Summary: Continuous $\times$ binary

Example: Y: height, pulse rate and X: gender, Treatment (only two: placebo and new treatment)
$\hookrightarrow$ two populations defined by the binary variable.

1. Comparison of two means:
   test or confidence interval for the population mean μ
   by T-test, Z-test (Normal distribution)
2. Linear model with Y as response variable and X as covariate.
   T-test on the slope

Summary: Continuous $\times$ binary

Example: Y: height, pulse rate and X: gender, Treatment (only two: placebo and new treatment)
$\hookrightarrow$ two populations defined by the binary variable.

1. Comparison of two means:
   test or confidence interval for the population mean μ
   by T-test, Z-test (Normal distribution)
2. Linear model with Y as response variable and X as covariate.
   T-test on the slope
3. Anova method with one factor (X and df=1)
   F-test

Summary: Continuous $\times$ binary

Example: Y: height, pulse rate and X: gender, Treatment (only two: placebo and new treatment)
$\hookrightarrow$ two populations defined by the binary variable.

1. Comparison of two means:
   test or confidence interval for the population mean μ
   by T-test, Z-test (Normal distribution)
2. Linear model with Y as response variable and X as covariate.
   T-test on the slope
3. Anova method with one factor (X and df=1)
   F-test

Summary: Continuous $\times$ categorical

Example: Y: height, pulse rate and X: Ethnicity, Treatment (several treatments)

Summary: Continuous $\times$ categorical

Example: Y: height, pulse rate and X: Ethnicity, Treatment (several treatments)

$\hookrightarrow$ several populations defined by the categorical variable.

Summary: Continuous $\times$ categorical

Example: Y: height, pulse rate and X: Ethnicity, Treatment (several treatments)

$\hookrightarrow$ several populations defined by the categorical variable.

1. Comparison of several means: Anova method with one factor (X and $df = C - 1$)
   F-test

Summary: Continuous $\times$ categorical

Example: Y: height, pulse rate and X: Ethnicity, Treatment (several treatments)
$\hookrightarrow$ several populations defined by the categorical variable.

1. Comparison of several means: Anova method with one factor (X and $df = C - 1$)
   F-test
2. Linear model with Y as response variable and X as covariate.

Summary: Continuous $\times$ Continuous

Example: Y: Weight, pulse rate and X: Height, rate of cholesterol)

Summary: Continuous $\times$ Continuous

Example: Y: Weight, pulse rate and X: Height, rate of cholesterol)

1. Correlation coefficient ρ

Summary: Continuous $\times$ Continuous

Example: Y: Weight, pulse rate and X: Height, rate of cholesterol)

1. Correlation coefficient ρ t-test or confidence interval

Summary: Continuous $\times$ Continuous

Example: Y: Weight, pulse rate and X: Height, rate of cholesterol)

1. Correlation coefficient ρ t-test or confidence interval
2. Linear model T-test on the slope

Summary: Continuous $\times$ Continuous

Example: Y: Weight, pulse rate and X: Height, rate of cholesterol)

1. Correlation coefficient ρ t-test or confidence interval
2. Linear model T-test on the slope
3. Anova method with one factor (X and df=1) F-test

Summary: Continuous $\times$ Continuous

Example: Y: Weight, pulse rate and X: Height, rate of cholesterol)

1. Correlation coefficient ρ t-test or confidence interval
2. Linear model T-test on the slope
3. Anova method with one factor (X and df=1) F-test

Summary: binary $\times$ binary

Example: Y: Disease Status X: gender, treatment (Placebo and one treatment)

Summary: binary $\times$ binary

Example: Y: Disease Status X: gender, treatment (Placebo and one treatment)
$\hookrightarrow$ two populations defined by the binary variable X.

1. Comparison on two proportions

Summary: binary $\times$ binary

Example: Y: Disease Status X: gender, treatment (Placebo and one treatment)
$\hookrightarrow$ two populations defined by the binary variable X.

1. Comparison on two proportions Z-test or confidence interval

Summary: binary $\times$ binary

Example: Y: Disease Status X: gender, treatment (Placebo and one treatment)
$\hookrightarrow$ two populations defined by the binary variable X.

1. Comparison on two proportions Z-test or confidence interval
2. Logistic model Z-test on the coefficient associated to the X variable

Summary: binary $\times$ binary

Example: Y: Disease Status X: gender, treatment (Placebo and one treatment)
$\hookrightarrow$ two populations defined by the binary variable X.

1. Comparison on two proportions Z-test or confidence interval
2. Logistic model Z-test on the coefficient associated to the X variable
3. Odds Ratio: Confidence interval

Summary: binary $\times$ binary

Example: Y: Disease Status X: gender, treatment (Placebo and one treatment)

$\hookrightarrow$ two populations defined by the binary variable X.

1. Comparison on two proportions Z-test or confidence interval
2. Logistic model Z-test on the coefficient associated to the X variable
3. Odds Ratio: Confidence interval
4. χ^2 for independence

Summary: binary $\times$ categorical

Example: Y: Status disease X: Ethnicity, treatment (several treatments)

Summary: binary $\times$ categorical

Example: Y: Status disease X: Ethnicity, treatment (several treatments)

	treatment_1	treatment_2	treatment_3	Total
Disease	a	b	c	
no Disease	d	e	f	
Total				

Summary: binary $\times$ categorical

Example: Y: Status disease X: Ethnicity, treatment (several treatments)

	treatment$_1$	treatment$_2$	treatment$_3$	Total
Disease	a	b	c	
no Disease	d	e	f	
Total				

1. χ^2 for independence

GOOD LUCK FOR YOUR EXAM

- Reading: 10 minutes
- Duration: 120 minutes
- Format: Short answer, Short essay, Problem solving
- Task Description:
  The final examination will cover the second half of the course (Chapters 13-25 in the textbook).
  In the exam you will be provided with statistical tables and a sheet of useful formulas. You will be permitted to bring a single double-sided A4 sheet of handwritten notes into the exam. (Photocopies of handwritten sheets are not permitted.)
  You should clearly write your name and student number on your A4 sheet - it will be collected at the end of the exam.