MA930 Data Analysis & Machine Learning

Lecture 3:Frequentist Statistics

Haoran Ni

THE UNIVERSITY OF WARWICK

Recap from last week

- Common distributions: e.g. Bernoulli, Binomial, Poisson, Normal, Gamma
- Characteristic functions $\phi_X(t) = \mathbf{E}(\exp(itX))$
- Statistics of a sample $X_1, X_2, X_3, \dots, X_n$ from a population with mean μ and variance σ^2
  - Sample mean: $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$, where $\mathbf{E}(\bar{X}) = \mu$
  - (Unbiased) sample variance: $s_u^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$. Hereafter we will simply refer to this as the "sample variance". $\mathbf{E}(s_u^2) = \sigma^2$.
- Law of large numbers. $\bar{X}_n \rightarrow \mu$ for many samples n.
- Central Limit Theorem: Distribution of $\bar{X}_n$ tends to a normal with mean μ and variance σ^2/n.

N.b. $\sigma/n^{1/2}$ is called the ”standard error on the mean”.

Outline

- Confidence intervals
- Hypothesis testing
- Common test

Motivation – confidence intervals

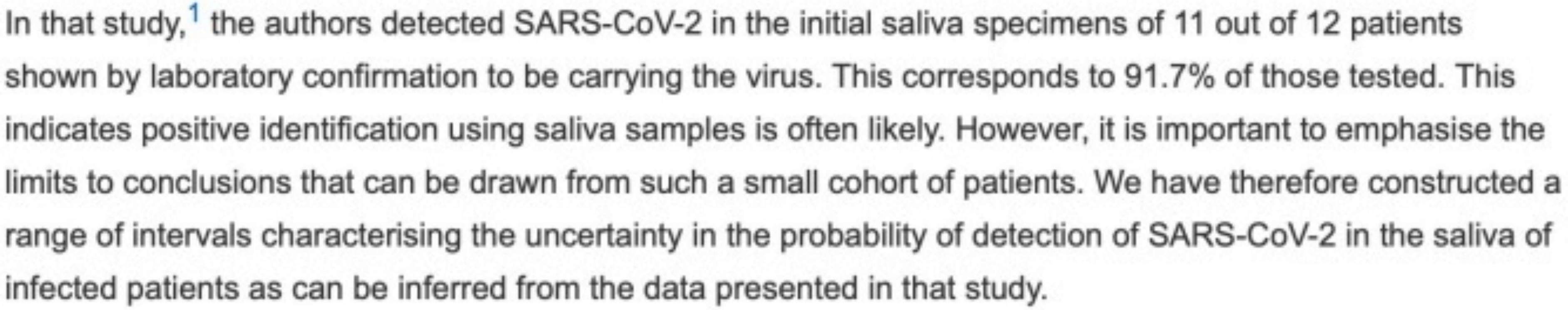
Example 1. Analysis of house price data.

In that example, there were 763,295 properties with price < £2M
Population mean = £305k, STD = £239k (highly skewed distribution)

Suppose that we take a sample of size n = 50. How can we estimate a range that the population mean might be in?

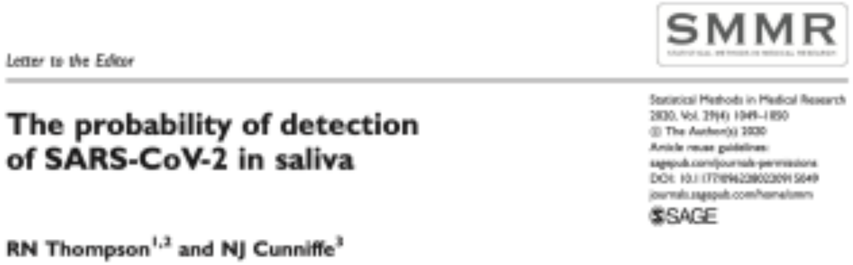
Motivation – confidence intervals

Example 2. COVID-19 saliva tests (early 2020).



In that study,[1] the authors detected SARS-CoV-2 in the initial saliva specimens of 11 out of 12 patients shown by laboratory confirmation to be carrying the virus. This corresponds to 91.7% of those tested. This indicates positive identification using saliva samples is often likely. However, it is important to emphasise the limits to conclusions that can be drawn from such a small cohort of patients. We have therefore constructed a range of intervals characterising the uncertainty in the probability of detection of SARS-CoV-2 in the saliva of infected patients as can be inferred from the data presented in that study.

What is the range that the true detection probability lies in?



SMMR

Letter to the Editor

The probability of detection of SARS-CoV-2 in saliva

Statistical Methods in Medical Research
2020, Vol. 29(4) 1049–1050
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280220915049
journals.sagepub.com/home/smm
SAGE

RN Thompson[1,2] and NJ Cunniffe[3]

Frequentist vs Bayesian

Two main approaches for thinking about statistical inference

Frequentist	Bayesian
Parameter values treated as *fixed* quantities (i.e. true value exists but is unknown, and we wish to estimate it)	Parameters are treated as random variables (i.e. they follow a probability distribution)
The frequentist interpretation of probability corresponds to the uncertainty in an outcome of an experiment each time, if the experiment was repeated an infinite number of times	The Bayesian interpretation of probability corresponds to how a degree of belief in a proposition changes due to available evidence
Confidence intervals (parameters fixed; bounds are random variables)	**Credible** intervals (bounds fixed; parameters are random variables)

Frequentist vs Bayesian

Two main approaches for thinking about statistical inference

Frequentist	Bayesian
Parameter values treated as *fixed* quantities (i.e. true value exists but is unknown, and we wish to estimate it)	Parameters are treated as random variables (i.e. they follow a probability distribution)
The frequentist interpretation of probability corresponds to the uncertainty in an outcome of an experiment each time, if the experiment was repeated an infinite number of times	The Bayesian interpretation of probability corresponds to how a degree of belief in a proposition changes due to available evidence
Confidence intervals (parameters fixed; bounds are random variables)	**Credible** intervals (bounds fixed; parameters are random variables)

We will cover some methods for inference under both approaches. Today, we will focus on frequentist approaches

Confidence intervals

- Confidence intervals are random quantities (i.e. they depend on a sample).
- Often misunderstood! A 95% confidence interval means that, if the experiment is repeated forever, 95% of confidence intervals will include the population mean

- In frequentist statistics, the population mean is **not** considered a random variable that has some probability of taking a particular value

N.b. confidence intervals are not unique

Confidence intervals

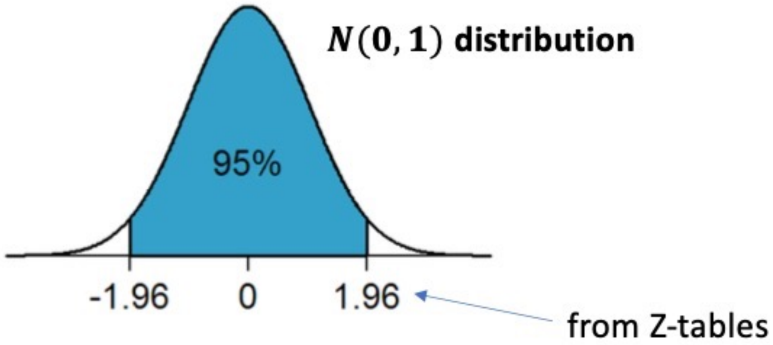
Example 3.1. Mean house prices (normal central confidence interval)

Suppose we sample n = 50 house prices, and that the sample mean μ_s= 280k and (unbiased) sample STD σ_s = 203k.

How do we estimate a range that the population mean is likely to be in?

By CLT, $\bar{X}$ is drawn from a $N(\mu, \frac{\sigma^2}{n})$ distribution so long as n is large enough

So $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ follows a $N(0,1)$ distribution



Confidence intervals

Example 3.1. Mean house prices (normal central confidence interval)

Suppose we sample n = 50 house prices, and that the sample mean μ_s= 280k and (unbiased) sample STD σ_s = 203k.

How do we estimate a range that the population mean is likely to be in?

By CLT, $\bar{X}$ is drawn from a $N(\mu, \frac{\sigma^2}{n})$ distribution so long as n is large enough

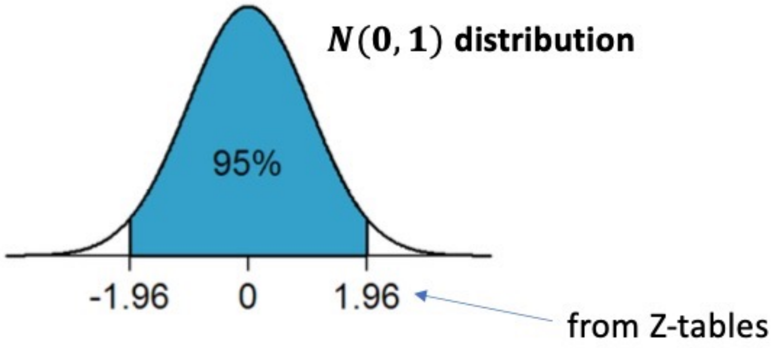
Confidence intervals

Example 3.1. Mean house prices (normal central confidence interval)

Suppose we sample n = 50 house prices, and that the sample mean μ_s= 280k and (unbiased) sample STD σ_s = 203k.

How do we estimate a range that the population mean is likely to be in?

$$P\left(-1.96 \leq \frac{\bar{X}-\mu}{\sigma\,/\sqrt{n}} \leq 1.96\right) = 0.95$$



Confidence intervals

Example 3.1. Mean house prices (normal central confidence interval)

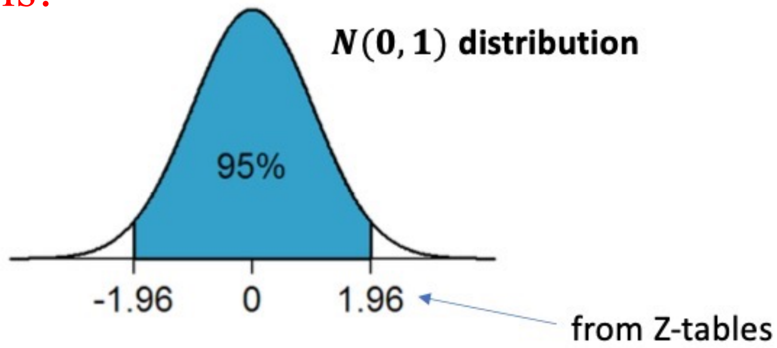
Suppose we sample n = 50 house prices, and that the sample mean μ_s= 280k and (unbiased) sample STD σ_s = 203k.

How do we estimate a range that the population mean is likely to be in?

$$P\left(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

$$P\left(\bar{X} - \frac{1.96\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right) = 0.95$$

Check this!



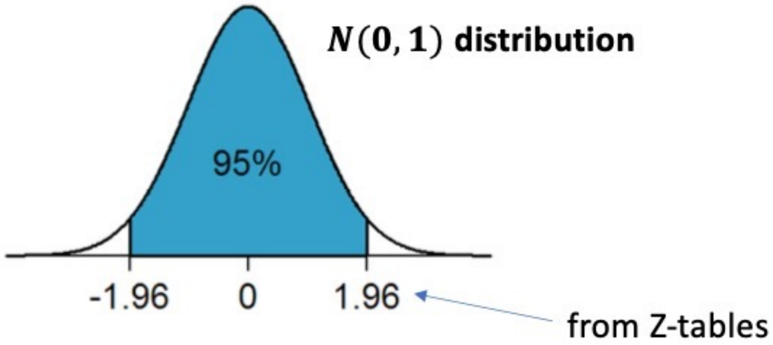
Confidence intervals

Example 3.1. Mean house prices (normal central confidence interval)

Suppose we sample n = 50 house prices, and that the sample mean μ_s= 280k and (unbiased) sample STD σ_s = 203k.

How do we estimate a range that the population mean is likely to be in?

Using sample STD to estimate population value, then population mean μ likely to lie in range $[\bar{X} - \frac{1.96\sigma_s}{\sqrt{n}}, \bar{X} + \frac{1.96\sigma_s}{\sqrt{n}}] = [223730{,}336270]$ ← **95% CI**



Confidence intervals

Example 3.1. Mean house prices (normal central confidence interval)

Suppose we sample n = 50 house prices, and that the sample mean μ_s= 280k and (unbiased) sample STD σ_s = 203k.

How do we estimate a range that the population mean is likely to be in?

Interpretation: The population mean μ is fixed. The confidence interval ends are random. Randomness arises from the random sampling of the sample mean.

N.b. Strictly speaking, it's not correct to state that the population mean is in the interval with probability 0.95 - the population mean is either in there or it isn't!

Confidence intervals

Exercise 3.1. Normal central confidence interval

100 samples from a gold bar machine had weights with mean $\mu_s = 101\text{g}$ and sample STD $\sigma_s = 5\text{g}$. Construct 99% confidence limits for the mean weight of all bars made.

HINT: look up a standard normal probability calculator online

Confidence intervals

Clopper-Pearson 95% confidence interval

For observations of binomially distributed variables, the Clopper-Pearson interval can be calculated to obtain a confidence interval on the success probability, p. To do this, suppose that we observe x successes out of n trials.

- Pick p_1 so that $P(X \geq x|p_1) = 0.025$ [i.e. you are unlikely to see so many successes if p were any smaller]
- Pick p_2 so that $P(X \leq x|p_2) = 0.025$ [i.e. you are unlikely to see so few successes if p were any bigger]

Then, the 95% Clopper-Pearson interval for p is $[p_1, p_2]$

Confidence intervals

Exercise 3.2. Clopper-Pearson 95% confidence interval

Write computing code to find the CI for:

i) A scenario in which 10 successes occur out of 20 trials

ii) Motivatory example 2 from the beginning of this lecture. Can we be confident from that evidence that saliva tests are highly effective at identifying SARS-CoV-2?

Outline

- Confidence intervals
- Hypothesis testing
- Common test

Hypothesis testing

Type I and Type II errors

Want to distinguish between a null hypothesis H_0 and other hypothesis H_1

In frequentist statistics, make a decision about the null hypothesis based on available Evidence

Four possibilities: Null is true/false, and we can/cannot reject it

	True	False
Do not reject	Success	Type II error
Reject	Type I error	Success

Probability of type I error $= \boldsymbol{P}$ (reject $H_0 |\ H_0$ true)

also called “type I error rate”, or $\boldsymbol{\alpha}$

Probability of type II error $= \boldsymbol{P}$(do not reject $H_0 |\ H_0$ false)

also called “type II error rate”, or $\boldsymbol{\beta}$

Hypothesis testing

Example 3.2 (board). Testing for disease

100 patients are tested for a disease using a new rapid test
The null hypothesis is that patients are healthy
Health or disease is then confirmed clinically later

What are the type-I and type-II error probabilities?

	Healthy	Disease
Test predicts healthy	88	4
Test predicts disease	2	6

Hypothesis testing

Example 3.2 (board). Testing for disease

100 patients are tested for a disease using a new rapid test
The null hypothesis is that patients are healthy
Health or disease is then confirmed clinically later

What are the type-I and type-II error probabilities?

	Healthy	Disease
Test predicts healthy	88	4
Test predicts disease	2	6

Note, these are not the same as quantities that sound similar e.g.

$\boldsymbol{P}(\text{healthy}|\text{test predicts disease})$, $\boldsymbol{P}(\text{disease}|\text{test predicts healthy})$

Frequentist ~ P(data | null hypothesis), Bayesian ~ P(hypothesis | data)

Significance and p-values

p-value: Probability you get the results you found or something more extreme, given the null hypothesis is true

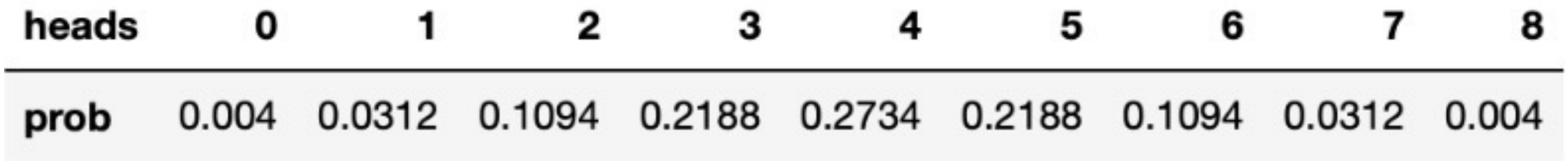
For example, imagine testing whether a coin is fair – this is the null hypothesis. Do an experiment with n flips.

Fix ahead a probability α (the ”significance” – type I error rate). This is the level below which we reject the null because result is unlikely if null true. Typically set at 5% or 1%. What is the p-value? Is it greater or less than α?

Significance and p-values

p-value: Probability you get the results you found or something more extreme, given the null hypothesis is true

Suppose experiment gives 7 heads out of 8 flips:



heads	0	1	2	3	4	5	6	7	8
prob	0.004	0.0312	0.1094	0.2188	0.2734	0.2188	0.1094	0.0312	0.004

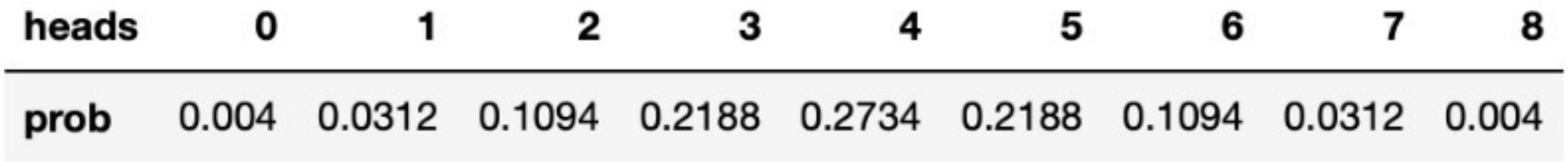
At least as extreme is 0, 1, 7 or 8 heads. P-value = 0.004 + 0.0312 + 0.0312 + 0.004, i.e. 7%.

There is *not significant* evidence to reject the null hypothesis at the 5% level

Significance and p-values

p-value: Probability you get the results you found or something more extreme, given the null hypothesis is true

If instead we test whether or not the coin is biased towards heads. H_0: Coin is not biased towards heads.



heads	0	1	2	3	4	5	6	7	8
prob	0.004	0.0312	0.1094	0.2188	0.2734	0.2188	0.1094	0.0312	0.004

At least as extreme is 7 or 8 heads. p-value = 0.0312 + 0.004, i.e. 3.5%. There is *significant* evidence to reject the null hypothesis at the 5% level.

First example was a two-sided test

This is a one-sided test

Significance and p-values

Exercise 3.3. Significance and p-values

A nationwide school test has mean mark $\mu = 75$ and standard deviation $\sigma = 7$. A particular school with 30 children has mean mark $\mu_S = 72$.

Write computing code to assess whether or not the school is significantly different from the national average.

Significance and p-values

Exercise 3.4.
Suppose that we toss a coin 100 times in order to determine whether or not the coin is fair, and that we decide the coin is fair if there are between 40-60 heads (inclusive).

i) What is the type I error probability?

ii) Write computing code to plot the probability of a type II error as a function of the probability each coin flip is a head (*p,* ignoring *p* = 0.5) – this is called an “operator curve”. Also plot the probability of rejecting the null hypothesis as a function of *p*.

Outline

- Confidence intervals
- Hypothesis testing
- Common test

Some common tests

One-sample test

Recap:

- Consider the sample mean $\bar{X}$ of a set of *n* numbers
- Assume *n* is sufficiently large that the sample mean is normal
- Central Limit Theorem: Distribution of $\bar{X}_n$ tends to a normal with mean μ and variance σ^2/n
- Calculate unbiased sample variance, $s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$, which is an estimate of σ^2

Some common tests

One-sample test

Recap:

- Consider the sample mean $\bar{X}$ of a set of *n* numbers
- Assume *n* is sufficiently large that the sample mean is normal
- Central Limit Theorem: Distribution of $\bar{X}_n$ tends to a normal with mean μ and variance σ^2/n
- Calculate unbiased sample variance, $s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$, which is an estimate of σ^2

- Note in general, if $X \sim N(\mu, \sigma^2)$ then $\frac{X-\mu}{\sigma} \sim N(0, 1)$

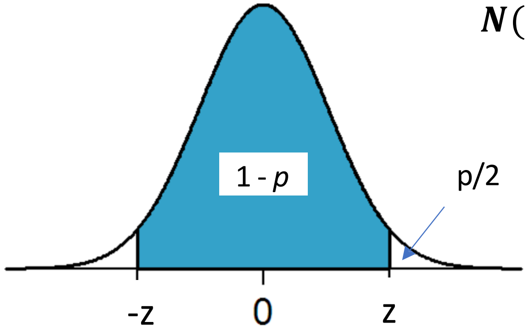
- So, here $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$

Some common tests

One-sample test

- Test whether or not the population mean takes a specific value. H$_0$: $\mu = \mu_0$
- Calculate z-statistic: $z = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$
- Calculate corresponding p-value (of the test statistic) in a normal distribution table (or using an online normal distribution calculator), and reject H$_0$ if z is too extreme (i.e. p is too small).

If null true…



$\boldsymbol{N(0,1)}$ **distribution (characterising** $\frac{\bar{X}-\mu_0}{s/\sqrt{n}}$**)**

n.b. this is a “two-sided” example

Some common tests

Exercise 3.5. One-sample test (two-sided)

Write computing code to generate n uniform random numbers on [0,1]. Test the hypothesis that the population mean is 0.5. Try for a few values of n. For which values of n are you least likely to reject H_0?

Some common tests

Two-sample test

- Consider now having two sets of numbers $\{X_k\}$ and $\{Y_k\}$
- Suppose there are n_x and n_y of each, and the sample means are $\bar{X}$ and $\bar{Y}$
- The unbiased sample variances are $s_x^2 = \frac{1}{n_x - 1} \sum_{k=1}^{n} (X_k - \bar{X})^2$ and the analogous expression for y.

Some common tests

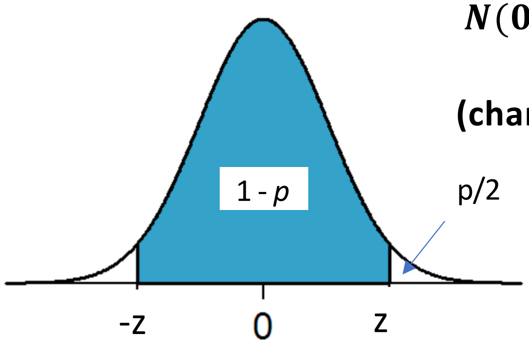
Two-sample test

- Are the X’s and Y’s drawn from distributions with the same mean? Consider statistics of $\bar{X} - \bar{Y}$
- By CLT: “the sample mean tends to a normal with mean μ and variance σ^2/n”
- Since the sum of normals is normal with summed means and variances,

the variance of the difference $(\bar{X} - \bar{Y})$ is $\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$

- The z statistic is now: $\frac{\bar{X} - \bar{Y} - 0}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$

If null true…



$N(0,1)$ distribution

(characterising $\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$)

Some common tests

Exercise 3.6. Two-sample test

In a test for intelligence, 20 MathSys students achieved the following scores: $\{65, 73, 51, 67, 48, 80, 69, 83, 89, 62, 71, 67, 64, 78, 85, 49, 80, 60, 51, 70\}$.

Students from another course achieved the following scores: $\{63, 72, 47, 63, 44, 78, 67, 52, 54, 58, 68, 65, 63, 77, 62, 46, 78, 56, 49, 65\}$.

Are students from both courses equally intelligent?

Some common tests

T-test

- Really, if n is small ($n < 30$), then distribution of X may not follow a normal distribution perfectly
- The z-score comes from a standardised normal: $\frac{\bar{X}-\mu}{s/\sqrt{n}}$

Some common tests

T-test

- Really, if n is small (n < 30), then distribution of X may not follow a normal distribution perfectly
- The z-score comes from a standardised normal: $\frac{\bar{X}-\mu}{s/\sqrt{n}}$
- If the population follows a normal distribution, then $\mathrm{t} = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ (follows a t-distribution with $n-1$ degrees of freedom (proof beyond the scope of this course)
- T-tests are then performed exactly like z-tests (n.b. the two-sample t-test is similar to the two sample z-test, but with the number of degrees of freedom in the t-distribution equal to $n_x - 1 + n_y - 1$)

Some common tests

Exercise 3.7. T-test

According to Public Health England, the mean height of adults in the UK is 66.5 inches. In a sample of 408 students from a sports club, the mean height was 67.4 inches, and the standard deviation was 10 inches.

- Use a t-test to investigate whether the height of individuals at the sports club is greater than the average for the UK population

- Repeat the analysis using a z-test, and check whether or not the conclusion is the same

Note that the Matlab command to return values from the cdf of the t- distribution is *tcdf*

Some common tests

Exercise 3.8. T-test

Load the “grades” data from the file that will be provided to you.

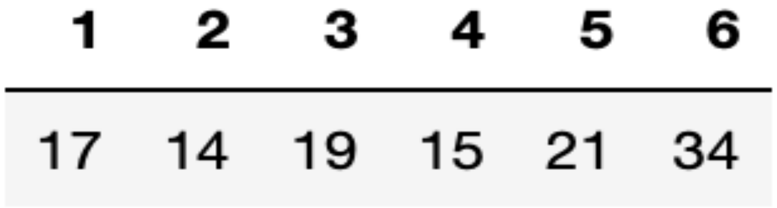
- Use a t-test to check whether or not the population mean test score is equal to $\mu_0 = 74$.
- Verify your conclusion using the Matlab command *ttest* (or equivalent in other coding languages.
- Write code to find the range of values of μ, for which the null hypothesis would not be rejected (and again verify your results using *ttest*).

Some common tests

Chi-square test

- Used for seeing if a distribution of numbers is as expected
- Suppose that $\{a_k\}_{k=1}^{n}$ is the measured frequency and $\{b_k\}_{k=1}^{n}$ is the expected frequency.
- The test statistics is $\chi^2 = \sum_{k=1}^{n} \frac{(a_k - b_k)^2}{b_k}$
- The sampling distribution is a chi-squared distribution with $n - 1$ degrees of freedom

Exercise 3.9. Chi-square test
A die is rolled 120 times, with the frequencies below.



1	2	3	4	5	6
17	14	19	15	21	34

Is the die a fair one? Test at the 95% and 99% levels, and give the p-value.

Additional Questions

Exercise 3.10. Spotting fake data

The first non-zero digit of numbers arising in real-world datasets do not follow a uniform distribution. There is a bias towards low numbers (Benford's Law):

$$\mathbf{P}(d) = \log_{10}\left(1 + \frac{1}{d}\right), \qquad \text{for } d = 1, 2, \dots, 9$$

- When financial data are faked, a uniform random number generator is often used.
- Consider the dataset of share prices. The third column is current stock prices (you will need to remove any "N/A"s or entries where the leading digit is zero).
- Plot the distribution of first digits.
- For a random sample of 45 stocks, (i) Plot the distribution of first digits; (ii) Check if the distribution of first digits is different from (a) Uniform; (b) Benford's Law. Give the p-values for these tests
- Increase the sample size to 450, and comment on the significance of the test in respect to Benford's Law.

Additional Questions

Exercise 3.11 (Optional). Limit of t-distribution

- The t-distribution is given by the pdf

$$f(t) = \frac{1}{\sqrt{v\pi}} \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})} \frac{1}{(1+\frac{t^2}{v})^{(v+1)/2}}$$

for $t \in [-\infty, \infty]$, where $v = n - 1$ is the number of degrees of freedom.

- Demonstrate that, in the limit of large v, this distribution approaches a standard normal.

Hint: Don't need to worry about the normalisation factor. Need to introduce an exponential to match a normal pdf. The Taylor expansion of log(1+x) may also be useful...

Additional Questions

Exercise 3.12 (Optional). One-sample test (one-sided)

1,500 people follow a particular diet for one month. A random sample of 29 people gained an average of 6.7 pounds, and the unbiased sample standard deviation was 7.1 pounds. Test the hypothesis that the average weight gain per person was more than 5 pounds.

Exercise 3.13 (Optional). Research exercise

Find an appropriate dataset online, and apply one of the tests described in this lecture.