1. Sampling

1.1 Sampling

census: every member of the population is observed.

Population: collection of individual items

Sample survey: small population of the population is observed.

Sampling units: individual units of a population

Sampling frame: a list of all sample

1.2 Using a random number table.

column为纵轴，从上往下走，如果digit不够要往后一个column走.
数字出现重复直接跳过

A factory makes safety ropes for climbers and has an order to supply 3000 ropes.
The buyer wants to know if the load at which the ropes break is more than a certain figure.
Suggest a reason why a census would not be used for this purpose.

the testing process will destroy all ropes

1.3 Random sampling

① simple random sampling / Label.

use a random number generator to select the required sample size in sampling frame.

ad: free of bias
inexpensive
each sampling unit has a known and equal chance of selection

dis:
Not suitable when population size is large
A sampling frame is needed

② Systematic sampling

randomly select a number between 1 - k
start with the __ having this __ number. then select the __ those have every kth __ number after that

$k = \frac{\text{population size}}{\text{sample size}}$

证明不是 systematic sampling.
sample is not being chosen at regular interval.

ad: suitable for large sample size.

dis: A sampling frame is needed
it introduces bias if sampling frame is not random.

③ Stratified sampling,

Label every strata.
每 strata 人数.

$\begin{cases} 1-a \\ 1-b \\ 1-c \end{cases}$

补充: How to improve reliability
Use larger sample
reduce bias → simple random sample.

ad: reflects population structure of ---

dis: population must be clearly classified

1.4 Non-random sampling

quota sampling:

① Divide the population into groups according to given characteristics.

前面步骤与 stratified sampling 一样.

② Once a quota has been filled, no more ___ are added.

Two improvements: increase the number of people asked

ask people at different times / locations

Chapter 2. Combinations of random variables

2.2 Combination of random variables

前提: random variables are normally distributed.

- If X and Y are two independent random variables, then:
  - $E(aX + bY) = aE(X) + bE(Y)$
  - $E(aX - bY) = aE(X) - bE(Y)$
- If X and Y are two independent random variables, then:
  - $Var(aX + bY) = a^2Var(X) + b^2Var(Y)$
  - $Var(aX - bY) = a^2Var(X) + b^2Var(Y)$

题型 $P(X>Y) \rightarrow X-Y$

$P(|Z|<1.44) = 2P(Z<1.44) - 1$

$P(|Y|>0.1) = 2P(Y>0.1)$

有名词 difference 一般就是绝对值.

Assumption: all random variables are independent. / sample are selected randomly.

Chapter 3. estimators and confidence intervals

For example, $\bar{X}$, the sample mean, is a statistic, whereas $\sum_{i=1}^{n}\frac{X_i^2}{n} - \mu^2$ is not a statistic since it involves the unknown population parameter μ.

- The sampling distribution of a statistic T is the probability distribution of T.

⑥ then X has distribution:

一般在设好 sample size 的时候

x	0	1
$P(X = x)$	$\frac{3}{5}$	$\frac{2}{5}$

$E(X) = \sum x P(X=x)$

$Var(X) = \sum x^2 P(X=x) - \mu^2$

- If a statistic T is used as an estimator for a population parameter θ and $E(T) = \theta$, then T is an unbiased estimator for θ.

已知.

A random sample $X_1, X_2, \ldots, X_n$ is taken from a population with $X \sim N(\mu, \sigma^2)$.

- An unbiased estimator for σ^2 is given by the sample variance S^2 where:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$= \frac{1}{n-1}\left(\sum x^2 - n\bar{x}^2\right)$$

- Standard error of $\bar{X} = \frac{\sigma}{\sqrt{n}}$ or $\frac{s}{\sqrt{n}}$

Twenty more days were randomly sampled. 给出 $\bar{x}$, s^2

已知 $s^2 = \frac{\sum y^2 - n\bar{x}^2}{n-1} \rightarrow$ 可得 $\sum y^2$ $\rightarrow$ 得出新的 $\sum x$ $\sum x^2$

已知 $\bar{x} = \frac{\sum x}{n} \rightarrow \sum x = n\bar{x}$.

If $X_i \sim N(\mu, \sigma^2)$ then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, where $\frac{\sigma}{\sqrt{n}}$ is the standard error.

For X_i follow binomial distribution.

$E(X) = np = \mu$ $Var(X) = np(1-p) = \sigma^2$

Standard error is a measure of the statistical accuracy of an estimator.

选哪个是 best estimator → 比较的 $Var(X)$ 小

3.2 Confidence intervals

- The 95% confidence interval for μ is $\left(\bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right)$

Notation The upper and lower values of a confidence interval are sometimes called the **confidence limits**.

(a, b)

已知 confidence interval $2\bar{x} = a + b$

size of ----- → 绝对值.

significance level 是拒绝域, confidence interval 是接受域

如果 X normally distribute, $\bar{x}$ normally distribute 是 exact 的

a Test, at the 5% level of significance, the doctor's claim. State your hypotheses clearly. (6

$\mu_A > \mu_B$

b State any assumptions you have made in testing the doctor's claim. (2

assumed normal distribution and that individual results are independent.

assumed $\sigma^2 = s^2$ for both population.

Chapter 4 Central limit theorem and testing the mean

4.1 The central limit theorem

This states that the mean of a large random sample taken from any random variable is always approximately normally distributed. This result is true without paying attention to the distribut of the original random variable.

accuracy 问题 → X 本身是否可以 normal distribution

sample size 是否 large.

4.2 Applying the central limit theorem to other distribution

Possion distribution. 遇到问 total number 时候 $\frac{\text{Total number}}{n}$ $P(X=a) = \frac{e^{-\lambda} \cdot \lambda^a}{a!}$

$\lambda = E(X)$ $\lambda = \sigma^2$ 先找 X 的分布，再找 sample size → $\bar{X}$

Problem-solving

If $\sum X_i$ is the sum of the observations from a sample of size n, then the sample mean is given by:

$\bar{X} = \frac{\sum X_i}{n}$

Binomial distribution.

$E(X) = np$ $Var(X) = np(1-p)$

uniformly distributed.

$E(X) = \frac{a+b}{2}$ $Var(X) = \frac{(b-a)^2}{12}$

4.3 Confidence intervals using the central limit theorem

importance of central limit theorem

Since the population is not normally distributed and n is large, we can use the central limit theorem to approximate the sample mean as a normal distribution.

c What is the relevance of the central limit theorem in finding this confidence interval?

b Since n is large, the central limit theorem allows us to approximate the mean distance travelle a normal distribution and so we can find a confidence interval for the mean distance travelled

b State whether or not it is necessary to assume that the value of merchandise sold has a normal distribution. Give a reason for your answer.

b It is not necessary to assume that the value of merchandise sold has a normal distribution because the sample size is large and we can use the central limit theorem.

4.4 Hypothesis testing the mean.

type $H_0: \mu = 60$ $H_1: \mu < 60$ 单边 双边 $H_0: \mu = 0.580$ $H_1: \mu \neq 0.580$

$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$

ex. $Z \leq -2.5758$ or $Z \geq 2.5758$

b Find the critical region for $\bar{X}$ in the above test.

The critical region of Z is:

$Z \leqslant -2.5758$ or $Z \geqslant 2.5758$

So $Z = \frac{\bar{X} - 0.580}{\frac{0.015}{\sqrt{50}}} \leqslant -2.5758$

4.5 Hypothesis testing for the difference between means

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

4.6 Use of large sample results for an unknown population

- If the population is normal, or can be assumed to be so, then, for large samples, $\frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$ has an approximate N(0, 1^2) distribution.

Watch out Both of these tests rely on large sample sizes, and the second test also relies on the central limit theorem.

- If the population is not normal, by assuming that s is a close approximation to σ, then for large samples, $\frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$ can be treated as having an approximate N(0, 1^2) distribution.

b State an assumption you have made in carrying out this test.

c Explain the significance of the central limit theorem to this test.

- The test statistic requires σ so you have to assume that $s^2 = \sigma^2$ for both samples.
- You are not told that the populations are normally distributed but the samples are both large and so the central limit theorem enables us to assume that $\overline{X}_A$ and $\overline{X}_B$ are both normal.

b State any assumptions you have made in testing the cardiologist's claim.

b Assume normal distribution or assume sample sizes large enough to use the central limit theorem; assume individual results are independent; assume $\sigma_2 = s_2$ for both populations.

Chapter 5 correlation

5.1 Spearman's rank correlation coefficient 一般diff positive correlation

strength of **linear correlation** between paired observations (x_i, y_i). In cases where the correlation is not linear, or where the data are not measurable on a **continuous** scale, the PMCC may not be a good measure of the correlation between two variables.

support the use of Spearman's

Spearman's rank correlation coefficient can be used instead of the product moment correlation coefficient if one of the following conditions is true:

one or both data sets are not from a normally distributed population

there is a non-linear relationship between the two data sets

one or both data sets already represent a ranking (as in Example 1).

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

When ranks are tied, the formula for the Spearman's rank correlation coefficient will not be correct and you will be required to use the PMCC formula instead.

Equal data values should be assigned a rank equal to the mean of the tied ranks.

Notation **Tied ranks** occur when two or more data values in one of the data sets are the same.

5.2 Hypothesis testing for zero correlation

① $H_0: \rho = 0$ $\begin{cases} H_1: \rho > 0 \\ H_2: \rho < 0 \end{cases}$

② $H_0: \rho = 0$ $H_2: \rho \neq 0$

③ critical region

1 The product moment correlation coefficient r is given by $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

$\sqrt{S_{xx}S_{yy}}$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 35$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 13($$

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

S_{xy}

to part **a** would change it:

i the literacy percentage for the eighth country was actually 77

ii a ninth country was added to the sample with life expectancy 79 years and literacy percentage 92%.

i The rank would still be the same, as the next highest percentage is 80. Therefore the coefficient would not change.

ii Both quantities would get the highest rank, thus $d = 0$. However, as n increases, the coefficient increases.

Chapter 6 Goodness of fit and contingency tables

6.1 Goodness of fit

用X^2来检验.

$$\boxed{X^2} = \sum \frac{(O_i - E_i)^2}{E_i}$$

Watch out The higher the value of X^2, the **less similar** the observed distribution is to the theoretical distribution.

H_0 H_1 一般是关于符不符合这个 distribution

6.2 Degrees of freedom and the χ^2 family of distribution

- Number of degrees of freedom = Number of cells (after any combining) − Number of constraints

- If any of the expected values are less than 5, then you have to combine expected frequencies in the data table until they are greater than 5.

Notation The χ^2 distribution is continuous, so $P(Y < y) = 1 - P(Y > y)$.

一般都是大于

Distribution	Degrees of freedom	
	Parameters known	Parameters not known
Discrete uniform	$n-1$	
Binomial	$n-1$	$n-2$
Poisson	$n-1$	$n-2$

Continuous uniform $n-1$

Normal $n-1$, then $n-2$ if one parameter not known, and $n-3$ if both parameters not known

6.3 Testing a hypothesis

Watch out A hypothesis test for goodness of fit is always **one-tailed**. This means the **critical region** is always the set of values **greater than** the critical value.

6.4 Testing the goodness of fit with discrete data.

① discrete uniform distribution.

Could the digits be from a random number table? Test at the 5% significance level.

Each digit should have an equal chance of selection, so the appropriate model is the discrete uniform distribution.

② binomial distribution

The conditions under which a binomial distribution arises are:

- there must be a fixed number (n) of trials in each observation
- the trials must be independent
- the trials have only two outcomes: success and failure
- the probability of success (p) is constant

$$p = \frac{\text{total number of successes}}{\text{number of trials} \times N} = \frac{\Sigma(r \times f_r)}{n \times N}$$

③ Poisson distribution

The conditions under which a Poisson distribution is likely to arise are:

- the events occur independently of each other
- the events occur singly and at random in continuous space or time
- the events occur at a constant rate, in the sense that the mean number in an interval is proportional to the length of the interval
- the mean and the variance are equal

Total number of observations $= N = \frac{8 \times 60}{6}$

$= 80$

$$\lambda = \frac{\Sigma(r \times f_r)}{N} = \frac{176}{80} = 2.2$$

6.5 Testing the goodness of fit with continuous data.

continuous uniform distribution

Class a to b	$b - a$	$\mathrm{P}(a < X < b)$ $\frac{b-a}{360 - 0}$	Frequency $\mathrm{P}(a < X < b) \times n$ $\mathrm{P}(a < X < b) \times 240$
$0 \leqslant d < 58$	58	0.1661	38.67
$58 \leqslant d < 100$	42	0.1167	28
$100 \leqslant d < 127$	27	0.075	18
$127 \leqslant d < 190$	63	0.175	42
$190 \leqslant d < 256$	66	0.1833	44
$256 \leqslant d < 296$	40	0.1111	26.66
$296 \leqslant d < 360$	64	0.1778	42.67

testing a normal distribution as a model.

Height (cm)	150-154	155-159	160-164	165-169	170-174	175-179	180-184	185-189	190-194
Frequency	4	6	12	30	64	52	18	10	4

Class
$X < 154.5$
$154.5 \leqslant X < 159.5$
$159.5 \leqslant X < 164.5$
$164.5 \leqslant X < 169.5$
$169.5 \leqslant X < 174.5$
$174.5 \leqslant X < 179.5$
$179.5 \leqslant X < 184.5$
$184.5 \leqslant X < 189.5$
$X \geqslant 189.5$

Describe how you would change this test if you were asked whether or not the height of male students could be modelled by a normal distribution with unknown mean and standard deviation.

We would now need to estimate the parameters:

	midpoints	frequency	fx	fx^2
150-154	152	4	608	92 416
155-159	157	6	942	147 894
160-164	162	12	1944	314 298
165-169	167	30	5010	836 670
170-174	172	64	11 008	1 893 376
175-179	177	52	9204	1 629 108
180-184	182	18	3276	596 232
185-189	187	10	1870	349 690
190-194	192	4	768	147 456

$\Sigma fx = 34\,630$

$\Sigma fx^2 = 6\,007\,770$

$n = 200$

$$\bar{x} = \frac{\Sigma fx}{n} = \frac{34\,630}{200} = 173.15$$

$$s^2 = \frac{1}{n-1}\left(\Sigma fx^2 - \frac{(\Sigma fx)^2}{n}\right) = \frac{1}{199}\left(6\,007\,770 - \frac{34\,630^2}{200}\right) = 58.22$$

6.6 Using contingency tables

We want to know if there is any association between the two schools' sets of results.

H_0: There is no association between the school and pass grade (school and pass grade are independent).

H_1: There is an association between school and pass grade (school and pass grade are not independent).

The expected frequencies (E_i) are:

	Salary				
	£0–£20k	£20k–£40k	£40–£60k	£60k–£80k	>£80k
Biology	2.90	67.29	26.40	4.51	2.90
Chemistry	3.01	69.88	27.42	4.68	3.01
Physics	3.09	71.82	28.18	4.81	3.09

Require each cell of the expected table to have a value at least 5; merge the first two columns (so create a category £0–£40K) and the last two columns (for a category >£60k).

	Salary		
	£0–£40	£40k–£60k	>£60k
Biology	70.19	26.40	7.41
Chemistry	72.89	27.42	7.69
Physics	74.92	28.18	7.90

The test statistic (X^2) calculations are: