HOLIDAY HOMEWORK 假期作业

AP Statistics

Exercise sheets and past-paper practice, topic by topic.
每个单元：练习卷 + 历年真题。

For class or self-study • no answer keys included.

Contents 目录

1 Describing one-variable data – Part 1

Name: ____________ Class: ________ Date: __________

Vocabulary 词汇

Write each word three times. 把每个单词抄写三遍。

English	中文	Write it 3 times (English)
mean	平均数	
median	中位数	
standard deviation	标准差	

Recall

You have worked with data before:

- You have found the average (mean) of a set of numbers.
- You have drawn bar charts and other graphs.
- Some data sets are more spread out than others.

This sheet lines up how to describe one set of data. Each idea has a worked example.

New ideas

1 The center

Two measures of the "middle" of data:

- the **mean** 平均数 – the average (add up, divide by how many),
- the **median** 中位数 – the middle value when the data are in order.

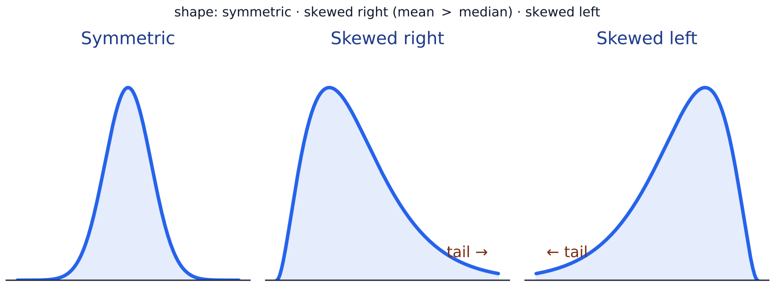
Worked example. For $2, 4, 4, 6, 9$: mean $= \dfrac{2+4+4+6+9}{5} = 5$; median $= 4$ (the middle one).

2 The spread

How spread out the data are: the **range** (largest minus smallest) and the **standard deviation** 标准差 (a typical distance from the mean).

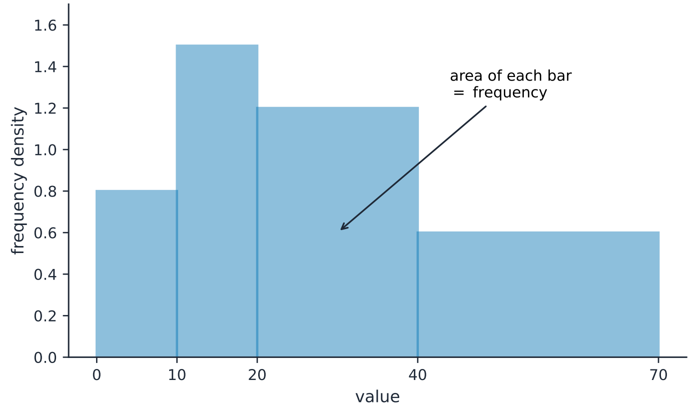
3 The shape

Data can be **symmetric** (even on both sides) or **skewed** (a long tail one way).



4 Pictures of data

A **histogram** shows the shape; a **box plot** shows the median and spread at a glance.



Watch out: the **mean** is pulled toward outliers (extreme values), but the **median** is not. For very skewed data (like incomes), the median often describes the "typical" value better.

Practice

Try these in order. The methods are all above.

2.1 State the difference between the mean and the median. [2]

2.2 Find the mean of $3, 5, 7, 9, 11$. [2]

2.3 Find the median of $3, 5, 7, 9, 11$. [1]

2.4 Find the range of $3, 5, 7, 9, 11$. [1]

2.5 A data set has one very large outlier. State whether the mean or the median is more affected by it, and why. [2]

STATISTICS

SECTION II

Part A

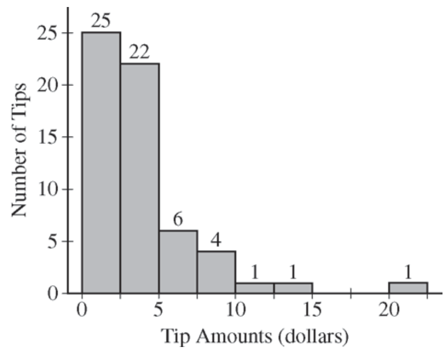
Questions 1-5

Spend about 65 minutes on this part of the exam.

Percent of Section II score—75

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

1. Robin works as a server in a small restaurant, where she can earn a tip (extra money) from each customer she serves. The histogram below shows the distribution of her 60 tip amounts for one day of work.



(a) Write a few sentences to describe the distribution of tip amounts for the day shown.

(b) One of the tip amounts was \$8. If the \$8 tip had been \$18, what effect would the increase have had on the following statistics? Justify your answers.

The mean:

The median:

4. As part of its twenty-fifth reunion celebration, the class of 1988 (students who graduated in 1988) at a state university held a reception on campus. In an informal survey, the director of alumni development asked 50 of the attendees about their incomes. The director computed the mean income of the 50 attendees to be $189,952. In a news release, the director announced, "The members of our class of 1988 enjoyed resounding success. Last year's mean income of its members was $189,952!"

   (a) What would be a statistical advantage of using the median of the reported incomes, rather than the mean, as the estimate of the typical income?

   (b) The director felt the members who attended the reception may be different from the class as a whole. A more detailed survey of the class was planned to find a better estimate of the income as well as other facts about the alumni. The staff developed two methods based on the available funds to carry out the survey.

   Method 1: Send out an e-mail to all 6,826 members of the class asking them to complete an online form. The staff estimates that at least 600 members will respond.

   Method 2: Select a simple random sample of members of the class and contact the selected members directly by phone. Follow up to ensure that all responses are obtained. Because method 2 will require more time than method 1, the staff estimates that only 100 members of the class could be contacted using method 2.

   Which of the two methods would you select for estimating the average yearly income of all 6,826 members of the class of 1988 ? Explain your reasoning by comparing the two methods and the effect of each method on the estimate.

1 The normal distribution – Part 2

Name: ____________ Class: ________ Date: __________

Vocabulary 词汇

Write each word three times. 把每个单词抄写三遍。

English	中文	Write it 3 times (English)
normal distribution	正态分布	

Recall

In Part A you described data's center, spread, and shape. Now one very common shape:

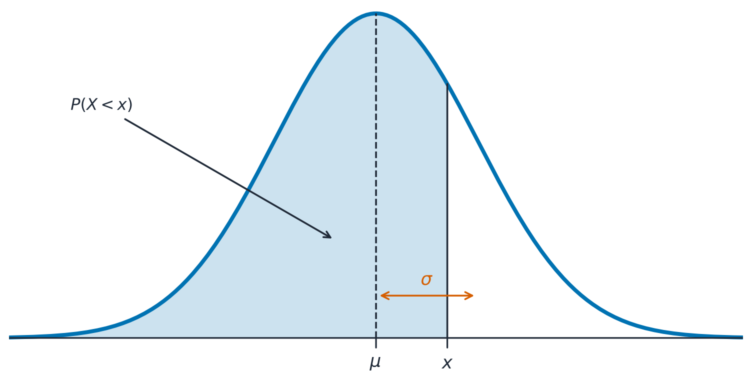
- Many things (heights, test scores) pile up in a "bell" shape.
- Most values are near the middle, with fewer far out.

Each idea has a worked example before the Practice.

New ideas

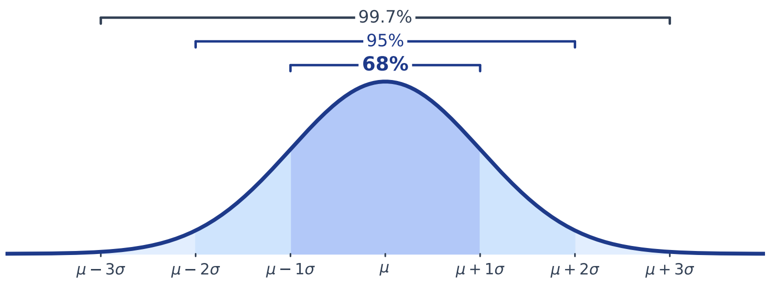
1 The normal distribution

A **normal distribution** 正态分布 is the classic **bell curve** – symmetric, with most data near the mean and less as you move away.



■ **2 The empirical rule**

For a normal distribution, the **empirical rule** (the 68–95–99.7 rule) says:



for approx. normal data: ~68% within 1σ, ~95% within 2σ, ~99.7% within 3σ

- about **68%** of values lie within 1 standard deviation of the mean,
- about **95%** within 2 standard deviations,
- about **99.7%** within 3.

■ **3 Using the rule**

Worked example. Test scores are normal with mean 100 and standard deviation 10. Then about 95% of scores lie between $100 - 2(10) = 80$ and $100 + 2(10) = 120$.

■ **4 Why it matters**

The bell curve appears so often that these percentages let you quickly judge how unusual a value is.

Watch out: the 68–95–99.7 rule only applies to a **normal** (bell-shaped) distribution. For skewed data it does not hold – always check the shape first.

Practice

Try these in order. The methods are all above.

2.1 Describe the shape of a normal distribution. [2]

2.2 State the three percentages of the empirical rule. [3]

2.3 Heights are normal with mean 170 cm and standard deviation 8 cm. State the range that holds about 68% of people. [2]

2.4 For the same heights, state the range that holds about 95%. [2]

2.5 State when the empirical rule does not apply. [1]

STATISTICS
SECTION II
Part A
Questions 1-5

Spend about 65 minutes on this part of the exam.

Percent of Section II score—75

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

1. Two large corporations, A and B, hire many new college graduates as accountants at entry-level positions. In 2009 the starting salary for an entry-level accountant position was $36,000 a year at both corporations. At each corporation, data were collected from 30 employees who were hired in 2009 as entry-level accountants and were still employed at the corporation five years later. The yearly salaries of the 60 employees in 2014 are summarized in the boxplots below.

(a) Write a few sentences comparing the distributions of the yearly salaries at the two corporations.

(b) Suppose both corporations offered you a job for $36,000 a year as an entry-level accountant.

(i) Based on the boxplots, give one reason why you might choose to accept the job at corporation A.

(ii) Based on the boxplots, give one reason why you might choose to accept the job at corporation B.

STATISTICS
SECTION II
Part A
Questions 1-5
Spend about 65 minutes on this part of the exam.
Percent of Section II score—75

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

1. Two large corporations, A and B, hire many new college graduates as accountants at entry-level positions. In 2009 the starting salary for an entry-level accountant position was $36,000 a year at both corporations. At each corporation, data were collected from 30 employees who were hired in 2009 as entry-level accountants and were still employed at the corporation five years later. The yearly salaries of the 60 employees in 2014 are summarized in the boxplots below.

(a) Write a few sentences comparing the distributions of the yearly salaries at the two corporations.

(b) Suppose both corporations offered you a job for $36,000 a year as an entry-level accountant.

(i) Based on the boxplots, give one reason why you might choose to accept the job at corporation A.

(ii) Based on the boxplots, give one reason why you might choose to accept the job at corporation B.

3. A grocery store purchases melons from two distributors, J and K. Distributor J provides melons from organic farms. The distribution of the diameters of the melons from Distributor J is approximately normal with mean 133 millimeters (mm) and standard deviation 5 mm.

   (a) For a melon selected at random from Distributor J, what is the probability that the melon will have a diameter greater than 137 mm?

   Distributor K provides melons from nonorganic farms. The probability is 0.8413 that a melon selected at random from Distributor K will have a diameter greater than 137 mm. For all the melons at the grocery store, 70 percent of the melons are provided by Distributor J and 30 percent are provided by Distributor K.

   (b) For a melon selected at random from the grocery store, what is the probability that the melon will have a diameter greater than 137 mm?

   (c) Given that a melon selected at random from the grocery store has a diameter greater than 137 mm, what is the probability that the melon will be from Distributor J?

2 Scatterplots and correlation – Part 1

Name: ____________ Class: ________ Date: __________

Vocabulary 词汇

Write each word three times. 把每个单词抄写三遍。

English	中文	Write it 3 times (English)
scatterplot	散点图	
Correlation	相关	

Recall

Often two things are linked:

- Taller people tend to weigh more.
- More study hours tend to mean higher marks.
- You can plot pairs of values to see a pattern.

This sheet lines up how to explore two variables together. Each idea has a worked example.

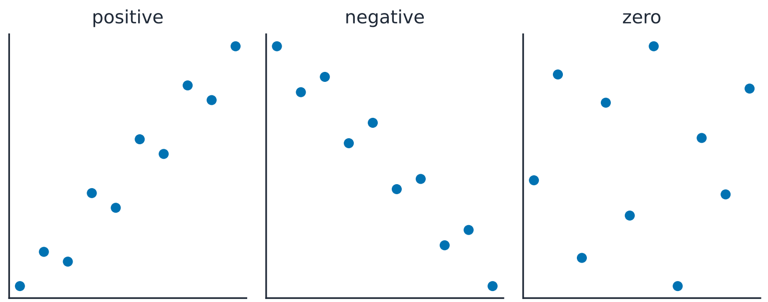
New ideas

1 Scatterplots

A **scatterplot** 散点图 plots each pair of values as a dot, so you can see any pattern between the two variables.

2 Correlation

Correlation 相关 describes the direction of a linear pattern:



- **positive** – as one rises, so does the other,
- **negative** – as one rises, the other falls,
- **none** – no clear pattern.

3 Strength

The correlation coefficient r measures how strong and which way the pattern is: near $+1$ is a strong positive line, near -1 a strong negative one, near 0 no linear link.

4 A caution

Worked example. Ice-cream sales and drownings both rise in summer – they are correlated, but neither causes the other. A third factor (hot weather) drives both.

Watch out: correlation is not causation. Two things rising together does not mean one causes the other – a hidden third factor may cause both.

Practice

Try these in order. The methods are all above.

2.1 State what a scatterplot shows. [2]

2.2 State what positive and negative correlation each mean. [2]

2.3 State what a correlation coefficient near +1 tells you. [2]

2.4 State what a correlation coefficient near 0 tells you. [1]

2.5 Ice-cream sales and sunburns are correlated. Explain why this does not mean ice cream causes sunburn. [2]

STATISTICS

SECTION II

Part A

Questions 1-5

Spend about 65 minutes on this part of the exam.

Percent of Section II score—75

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

1. Researchers studying a pack of gray wolves in North America collected data on the length x, in meters, from nose to tip of tail, and the weight y, in kilograms, of the wolves. A scatterplot of weight versus length revealed a relationship between the two variables described as positive, linear, and strong.

   (a) For the situation described above, explain what is meant by each of the following words.

   (i) Positive:

   (ii) Linear:

   (iii) Strong:

   The data collected from the wolves were used to create the least-squares equation $\hat{y} = -16.46 + 35.02x$.

   (b) Interpret the meaning of the slope of the least-squares regression line in context.

   (c) One wolf in the pack with a length of 1.4 meters had a residual of -9.67 kilograms. What was the weight of the wolf?

STATISTICS
SECTION II
Part A
Questions 1-5

Spend about 65 minutes on this part of the exam.

Percent of Section II score—75

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

1. An administrator at a large university is interested in determining whether the residential status of a student is associated with level of participation in extracurricular activities. Residential status is categorized as on campus for students living in university housing and off campus otherwise. A simple random sample of 100 students in the university was taken, and each student was asked the following two questions.

   - Are you an on campus student or an off campus student?
   - In how many extracurricular activities do you participate?

   The responses of the 100 students are summarized in the frequency table shown.

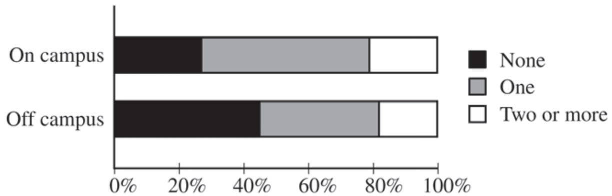
	Residential Status		
Level of Participation in Extracurricular Activities	On campus	Off campus	Total
No activities	9	30	39
One activity	17	25	42
Two or more activities	7	12	19
Total	33	67	100

(a) Calculate the proportion of on campus students in the sample who participate in <u>at least one</u> extracurricular activity and the proportion of off campus students in the sample who participate in <u>at least one</u> extracurricular activity.

On campus proportion:

Off campus proportion:

The responses of the 100 students are summarized in the segmented bar graph shown.



(b) Write a few sentences summarizing what the graph reveals about the association between residential status and level of participation in extracurricular activities among the 100 students in the sample.

(c) After verifying that the conditions for inference were satisfied, the administrator performed a chi-square test of the following hypotheses.

H_0 : There is no association between residential status and level of participation in extracurricular activities among the students at the university.

H_a : There is an association between residential status and level of participation in extracurricular activities among the students at the university.

The test resulted in a *p*-value of 0.23. Based on the *p*-value, what conclusion should the administrator make?

2. Product advertisers studied the effects of television ads on children's choices for two new snacks. The advertisers used two 30-second television ads in an experiment. One ad was for a new sugary snack called Choco-Zuties, and the other ad was for a new healthy snack called Apple-Zuties.

For the experiment, 75 children were randomly assigned to one of three groups, A, B, or C. Each child individually watched a 30-minute television program that was interrupted for 5 minutes of advertising. The advertising was the same for each group with the following exceptions.

- The advertising for group A included the Choco-Zuties ad but not the Apple-Zuties ad.
- The advertising for group B included the Apple-Zuties ad but not the Choco-Zuties ad.
- The advertising for group C included neither the Choco-Zuties ad nor the Apple-Zuties ad.

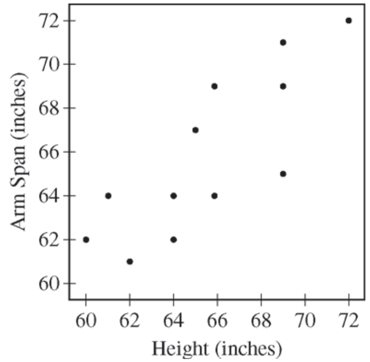
After the program, the children were offered a choice between the two snacks. The table below summarizes their choices.

Group	Type of Ad	Number Who Chose Choco-Zuties	Number Who Chose Apple-Zuties
A	Choco-Zuties only	21	4
B	Apple-Zuties only	13	12
C	Neither	22	3

(a) Do the data provide convincing statistical evidence that there is an association between type of ad and children's choice of snack among all children similar to those who participated in the experiment?

(b) Write a few sentences describing the effect of each ad on children's choice of snack.

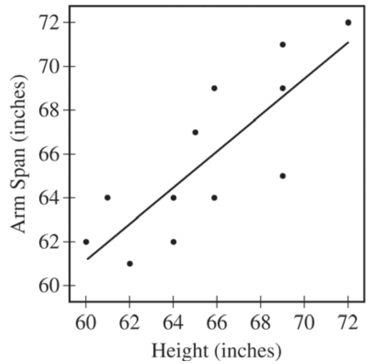
5. A student measured the heights and the arm spans, rounded to the nearest inch, of each person in a random sample of 12 seniors at a high school. A scatterplot of arm span versus height for the 12 seniors is shown.



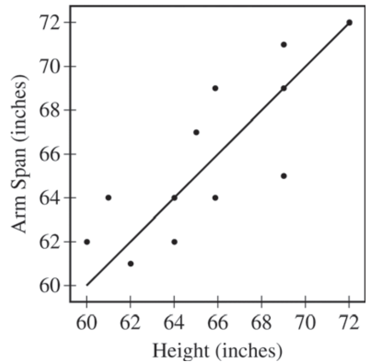
(a) Based on the scatterplot, describe the relationship between arm span and height for the sample of 12 seniors.

Let x represent height, in inches, and let y represent arm span, in inches. Two scatterplots of the same data are shown below. Graph 1 shows the data with the least squares regression line $\hat{y} = 11.74 + 0.8247x$, and graph 2 shows the data with the line $y = x$.

GRAPH 1



GRAPH 2



(b) The criteria described in the table below can be used to classify people into one of three body shape categories: square, tall rectangle, or short rectangle.

Square	Tall Rectangle	Short Rectangle
Arm span is equal to height.	Arm span is less than height.	Arm span is greater than height.

(i) For which graph, 1 or 2, is the line helpful in classifying a student's body shape as square, tall rectangle, or short rectangle? Explain.

(ii) Complete the table of classifications for the 12 seniors.

Classification	Square	Tall Rectangle	Short Rectangle
Frequency			

(c) Using the best model for prediction, calculate the predicted arm span for a senior with height 61 inches.

2 The line of best fit – Part 2

Name: ______ Class: ______ Date: ______

Vocabulary 词汇

Write each word three times. 把每个单词抄写三遍。

English	中文	Write it 3 times (English)
line of best fit	回归线	

Recall

In Part A you saw scatterplots and correlation. Now summarize the pattern with a line:

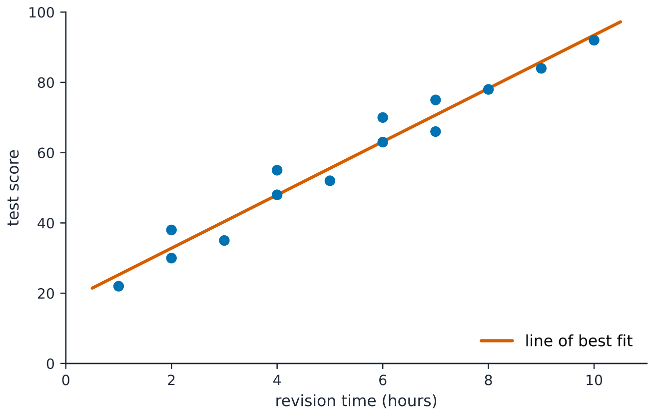
- A straight line can capture the trend in a scatterplot.
- That line lets you make predictions.

Each idea has a worked example before the Practice.

New ideas

1 The line of best fit

The **line of best fit** 回归线 (also called the regression line) is the straight line that best follows the trend of the points.



2 Reading the line

The line has the form $\hat{y} = a + bx$. The **slope** b says how much y changes for each 1 unit of x; the **intercept** a is the predicted y when $x = 0$.

3 Making predictions

Worked example. If $\hat{y} = 50 + 5x$ predicts a test score from hours studied, then 3 hours predicts $\hat{y} = 50 + 5(3) = 65$.

4 Residuals

A **residual** is the gap between an actual value and the line's prediction (actual minus predicted). Small residuals mean the line fits well.

Watch out: avoid **extrapolating** – predicting far outside the range of your data. The pattern may not continue, so a line that fits well for 1–5 study hours may give nonsense for 20 hours.

Practice

Try these in order. The methods are all above.

2.1 State what the line of best fit does. [2]

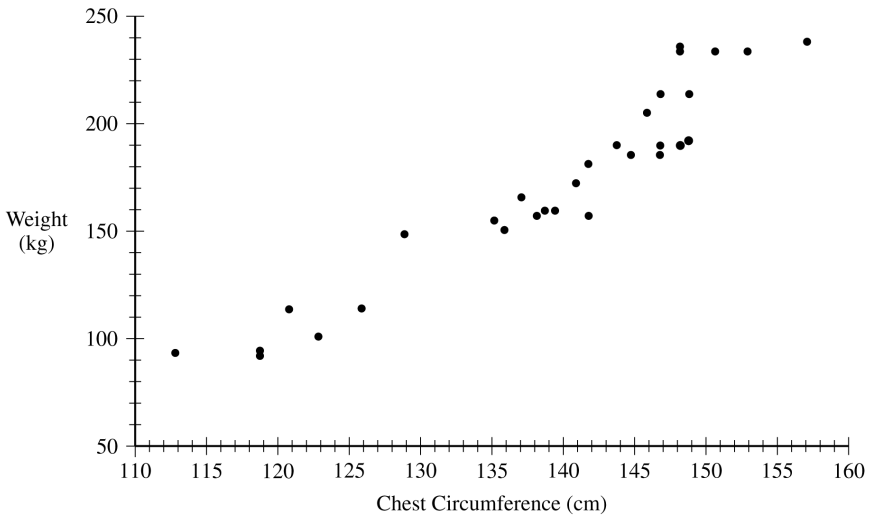
2.2 For $\hat{y} = 50 + 5x$, state what the slope 5 means. [2]

2.3 Using $\hat{y} = 50 + 5x$, predict y when $x = 4$. [2]

2.4 State what a residual is. [2]

2.5 Explain why predicting far outside the data range is risky. [2]

. Wildlife biologists are interested in the health of tule elk, a species of deer found in California. An important measurement of tule elk health is their weight. The weight of a tule elk is difficult to measure in the wild. However, chest circumference, which is believed to be related to the weight of a tule elk, can easily be measured from a safe distance using a harmless laser. A study was done to investigate whether chest circumference, in centimeters (cm), could be used to accurately estimate the weight, in kilograms (kg), of male tule elk. For the study, wildlife biologists captured 30 male tule elk, measured their chest circumference and weight, and then released the elk. The data for the 30 male tule elk are shown in the scatterplot.



(a) Describe the relationship between chest circumference and weight of male tule elk in context.

GO ON TO THE NEXT PAGE.

Continue your response to **QUESTION 5** on this page.

Following is the equation of the least-squares regression line relating chest circumference and weight for male tule elk.

$$\text{predicted weight} = -350.3 + 3.7455(\text{chest circumference})$$

(b) The weight of one male tule elk with a chest circumference of 145.9 cm is 204.3 kg.

(i) Using the equation of the least-squares regression line, calculate the predicted weight for this male tule elk. Show your work.

(ii) Calculate the residual for this male tule elk. Show your work.

GO ON TO THE NEXT PAGE.

Continue your response to **QUESTION 5** on this page.

The equation of the least-squares regression line relating chest circumference and weight for male tule elk is repeated here.

$$\text{predicted weight} = -350.3 + 3.7455(\text{chest circumference})$$

(c) Interpret the slope of the least-squares regression line in context.

(d) The sambar, another species of deer, is similar in size to the tule elk. The slope of the population regression line relating chest circumference and weight for all male sambars is 4.5 kilograms per centimeter. A wildlife biologist wants to determine whether the slope of the population regression line for male tule elk is different than that for male sambars. Let β represent the slope of the population regression line for male tule elk. The wildlife biologist conducted a test of the following hypotheses using the sample of 30 tule elk.

$$H_0: \beta = 4.5$$
$$H_a: \beta \neq 4.5$$

The test statistic was calculated to be 3.408. Assume all conditions for inference were met.

(i) Determine the p-value of the test.

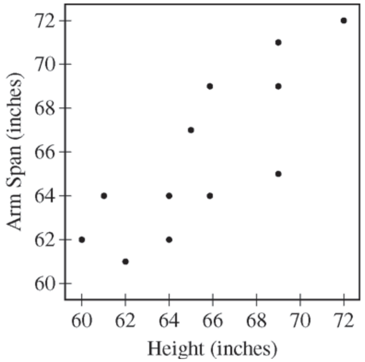
GO ON TO THE NEXT PAGE.

Continue your response to **QUESTION 5** on this page.

(ii) At a significance level of $\alpha = 0.05$, what conclusion should the wildlife biologist make regarding the slope of the population regression line for male tule elk? Justify your response.

GO ON TO THE NEXT PAGE.

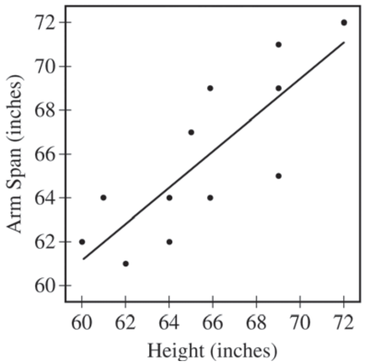
5. A student measured the heights and the arm spans, rounded to the nearest inch, of each person in a random sample of 12 seniors at a high school. A scatterplot of arm span versus height for the 12 seniors is shown.



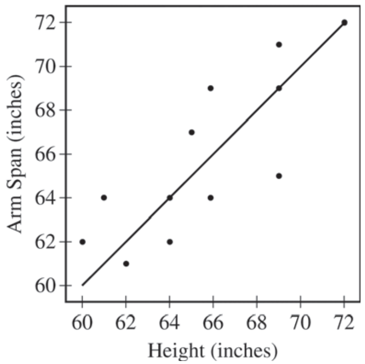
(a) Based on the scatterplot, describe the relationship between arm span and height for the sample of 12 seniors.

Let x represent height, in inches, and let y represent arm span, in inches. Two scatterplots of the same data are shown below. Graph 1 shows the data with the least squares regression line $\hat{y} = 11.74 + 0.8247x$, and graph 2 shows the data with the line $y = x$.

GRAPH 1



GRAPH 2



(b) The criteria described in the table below can be used to classify people into one of three body shape categories: square, tall rectangle, or short rectangle.

Square	Tall Rectangle	Short Rectangle
Arm span is equal to height.	Arm span is less than height.	Arm span is greater than height.

(i) For which graph, 1 or 2, is the line helpful in classifying a student's body shape as square, tall rectangle, or short rectangle? Explain.

(ii) Complete the table of classifications for the 12 seniors.

Classification	Square	Tall Rectangle	Short Rectangle
Frequency			

(c) Using the best model for prediction, calculate the predicted arm span for a senior with height 61 inches.

STATISTICS
SECTION II
Part A
Questions 1-5

Spend about 65 minutes on this part of the exam.

Percent of Section II score—75

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

1. Researchers studying a pack of gray wolves in North America collected data on the length x, in meters, from nose to tip of tail, and the weight y, in kilograms, of the wolves. A scatterplot of weight versus length revealed a relationship between the two variables described as positive, linear, and strong.

(a) For the situation described above, explain what is meant by each of the following words.

(i) Positive:

(ii) Linear:

(iii) Strong:

The data collected from the wolves were used to create the least-squares equation $\hat{y} = -16.46 + 35.02x$.

(b) Interpret the meaning of the slope of the least-squares regression line in context.

(c) One wolf in the pack with a length of 1.4 meters had a residual of -9.67 kilograms. What was the weight of the wolf?

Name: AP Statistics: 2017

STATISTICS
SECTION II
Part A
Questions 1-5

Spend about 65 minutes on this part of the exam.

Percent of Section II score—75

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

1. Researchers studying a pack of gray wolves in North America collected data on the length x, in meters, from nose to tip of tail, and the weight y, in kilograms, of the wolves. A scatterplot of weight versus length revealed a relationship between the two variables described as positive, linear, and strong.

(a) For the situation described above, explain what is meant by each of the following words.

(i) Positive:

(ii) Linear:

(iii) Strong:

The data collected from the wolves were used to create the least-squares equation $\hat{y} = -16.46 + 35.02x$.

(b) Interpret the meaning of the slope of the least-squares regression line in context.

(c) One wolf in the pack with a length of 1.4 meters had a residual of -9.67 kilograms. What was the weight of the wolf?

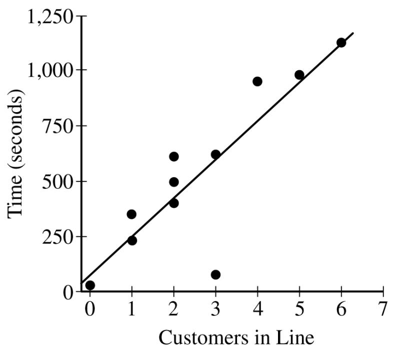
STATISTICS
SECTION II
Part A
Questions 1-5

Spend about 1 hour and 5 minutes on this part of the exam.

Percent of Section II score—75

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

1. The manager of a grocery store selected a random sample of 11 customers to investigate the relationship between the number of customers in a checkout line and the time to finish checkout. As soon as the selected customer entered the end of a checkout line, data were collected on the number of customers in line who were in front of the selected customer and the time, in seconds, until the selected customer was finished with the checkout. The data are shown in the following scatterplot along with the corresponding least-squares regression line and computer output.



Predictor	Coef	SE Coef	T	P
Constant	72.95	110.36	0.66	0.525
Customers in line	174.40	35.06	4.97	0.001
S = 200.01		R-Sq = 73.33%	R-Sq (adj) = 70.37%	

(a) Identify and interpret in context the estimate of the intercept for the least-squares regression line.

(b) Identify and interpret in context the coefficient of determination, r^2.

(c) One of the data points was determined to be an outlier. Circle the point on the scatterplot and explain why the point is considered an outlier.

3 Sampling – Part 1

Name: ____________ Class: ________ Date: __________

Vocabulary 词汇

Write each word three times. 把每个单词抄写三遍。

English	中文	Write it 3 times (English)
population	总体	
sample	样本	
random sample	随机抽样	
Bias	偏差	

Recall

To learn about a big group, you cannot ask everyone:

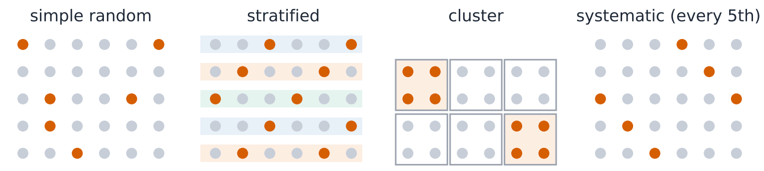
- To find the average height of a country, you cannot measure every person.
- Instead you measure a smaller group and use it to estimate.
- But the smaller group must be chosen fairly.

This sheet lines up how to collect data by sampling. Each idea has a worked example.

New ideas

1 Population and sample

The **population** 总体 is the whole group you want to study; a **sample** 样本 is the smaller part you actually measure.



SRS: equal chance · stratified: sample within groups · cluster: take whole clusters · systematic: every *k*th

2 Random sampling

A **random sample** 随机抽样 gives everyone an equal chance of being chosen. This keeps the sample fair and representative.

3 Bias

Bias 偏差 is anything that makes a sample unrepresentative – for example, only surveying people online misses those without internet.

4 A worked example

Worked example. To learn students' opinions, surveying only the football team is biased (not representative); drawing names randomly from the whole school is fair.

Watch out: a **biased** sample leads to wrong conclusions no matter how **large** it is. A huge but unfair sample is worse than a smaller, properly random one.

Practice

Try these in order. The methods are all above.

2.1 State the difference between a population and a sample. [2]

2.2 State what makes a sample "random". [2]

2.3 State what "bias" means in sampling. [2]

2.4 A survey about a town is done only outside a gym. Explain why this could be biased. [2]

2.5 State whether a large but biased sample or a smaller random sample is more trustworthy, and why. [2]

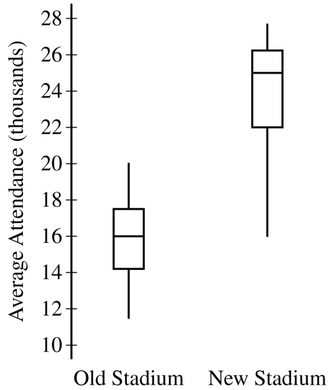
Part B

Question 6

Spend about 25 minutes on this part of the exam.

)irections: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness f your methods as well as on the accuracy and completeness of your results and explanations.

5. Attendance at games for a certain baseball team is being investigated by the team owner. The following boxplots summarize the attendance, measured as average number of attendees per game, for 47 years of the team's existence. The boxplots include the 30 years of games played in the old stadium and the 17 years played in the new stadium.

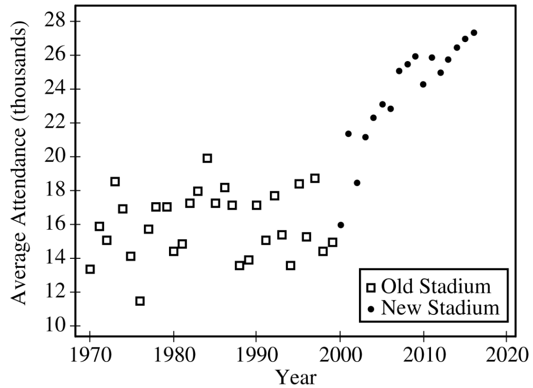


(a) Compare the distributions of average attendance between the old and new stadiums.

GO ON TO THE NEXT PAGE.

Use a pencil or pen with black or dark blue ink only. Do NOT write your name. Do NOT write outside the box.

The following scatterplot shows average attendance versus year.

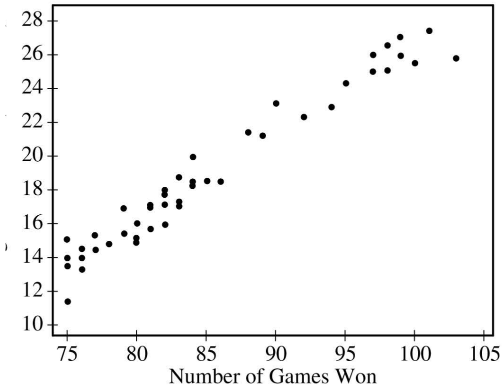


(b) Compare the trends in average attendance over time between the old and new stadium.

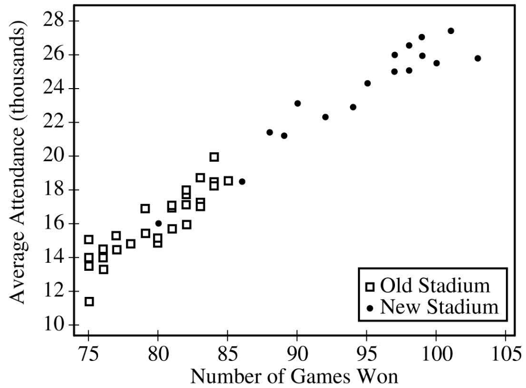
GO ON TO THE NEXT PAGE.

(c) Consider the following scatterplots.

Graph I



Graph II



(i) Graph I shows the average attendance versus number of games won for each year. Describe the relationship between the variables.

(ii) Graph II shows the same information as Graph I, but also indicates the old and new stadiums. Does Graph II suggest that the rate at which attendance changes as number of games won increases is different in the new stadium compared to the old stadium? Explain your reasoning.

GO ON TO THE NEXT PAGE.

(d) Consider the three variables: number of games won, year, and stadium. Based on the graphs, explain how one of those variables could be a confounding variable in the relationship between average attendance and the other variables.

GO ON TO THE NEXT PAGE.

STOP

END OF EXAM

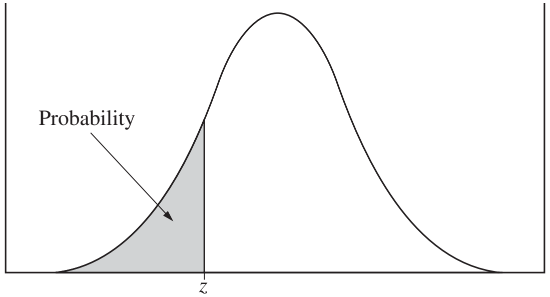


Table entry for z is the probability lying below z.

Table A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

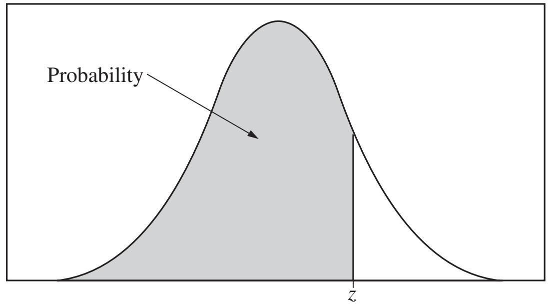


Table entry for z is the probability lying below z.

Table A (*Continued*)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Table entry for p and C is the point t^* with probability p lying above it and probability C lying between $-t^*$ and t^*.

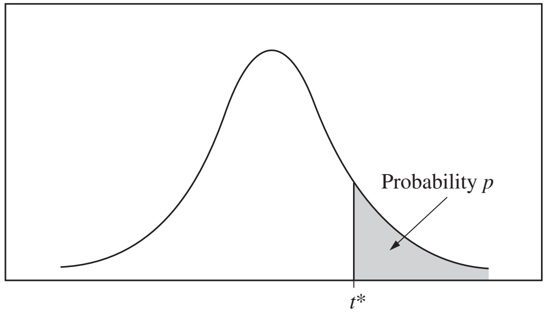


Table B ***t* distribution critical values**

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Table entry for p is the point (χ^2) with probability p lying above it.

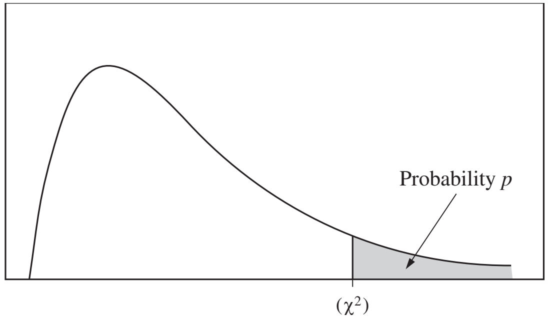


Table C χ^2 **critical values**

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.7
80	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8	128.3
100	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4	153.2

3. Alzheimer's disease results in a loss of cognitive ability beyond what is expected with typical aging. A local newspaper published an article with the following headline.

Study Finds Strong Association Between Smoking and Alzheimer's

The article reported that a study tracked the medical histories of 21,123 men and women for 23 years. The article stated that, for those who smoked at least two packs of cigarettes a day, the risk of developing Alzheimer's disease was 2.57 times the risk for those who did not smoke.

(a) Identify the explanatory and response variables in the study.

Explanatory variable:

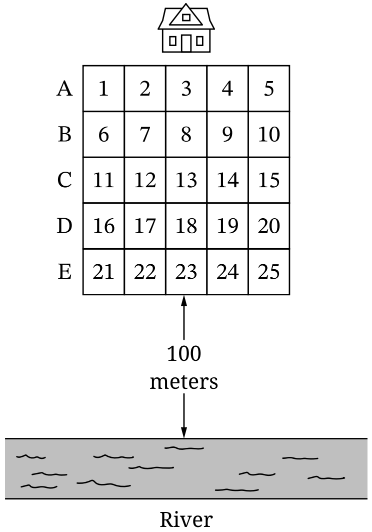
Response variable:

(b) Is the study described in the article an observational study or an experiment? Explain.

(c) Exercise status (regular weekly exercise versus no regular weekly exercise) was mentioned in the article as a possible confounding variable. Explain how exercise status could be a confounding variable in the study.

2. Aphids are tiny insects that feed on plants such as cabbage plants. A farmer wants to reduce the number of aphids in a cabbage field. A river is located 100 meters south of the cabbage field. The farmer divides the field into 25 regions of equal size, as shown in the diagram. Each region has approximately the same number of cabbage plants.

Farmer's House and Cabbage Field



The farmer would like to estimate the proportion of cabbage plants in the field that are affected by aphids and believes that the extent of aphid damage is greater for the regions in the cabbage field closer to the river. To obtain the estimate, the farmer is considering three sampling methods.

- Sampling method I: Select region 3, which is closest to the farmer's house and farthest from the river. Examine every cabbage plant in the region for aphid damage.
- Sampling method II: Randomly select one row (A, B, C, D, or E). For every region in the selected row, examine every cabbage plant for aphid damage.
- Sampling method III: Randomly select one region from each of rows A, B, C, D, and E. For each selected region, examine every cabbage plant for aphid damage.

A. Explain whether sampling method I is an appropriate sampling method for the farmer to use to estimate the proportion of cabbage plants in the field that are damaged by aphids.

B. Using sampling method II, the farmer randomly selected row E and examined every cabbage plant in row E. If the farmer's belief is correct, determine whether the selection of row E is likely to provide an overestimate or an underestimate of the proportion of cabbage plants in the field that are damaged by aphids. Justify your answer.

C. Using the information provided in the diagram of the cabbage field, describe how to implement sampling method III, which requires a random selection of one region from each of rows A, B, C, D, and E.

3 Experiments – Part 2

Name: ____________ Class: ________ Date: __________

Vocabulary 词汇

Write each word three times. 把每个单词抄写三遍。

English	中文	Write it 3 times (English)
experiment	实验	
control group	对照组	
randomization	随机化	

Recall

In Part A you met sampling. Now how to test whether something actually **causes** an effect:

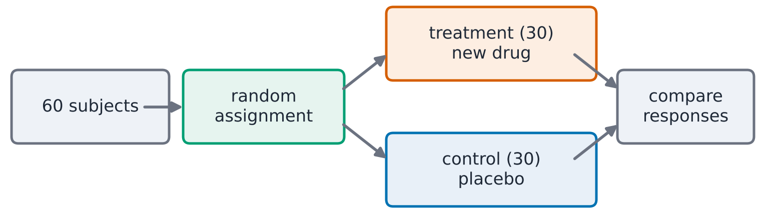
- Does a new medicine really work?
- Just observing is not enough – you need a fair test.

Each idea has a worked example before the Practice.

New ideas

1 Experiment versus observation

- An **observational study** just watches and records, without changing anything.
- An **experiment** 实验 deliberately applies a **treatment** to see its effect.



random assignment → causal comparison of responses

2 Control and treatment groups

An experiment compares a **treatment group** (gets the new thing) with a **control group** 对照组 (does not) – so any difference can be linked to the treatment.

3 Randomization

Subjects are put into groups by **randomization** 随机化, so the two groups are similar to begin with and the comparison is fair.

4 Only experiments show causation

Worked example. To test a medicine, randomly split patients: half get the drug, half get a placebo. Because the groups start alike, a difference in recovery can be blamed on the drug.

Watch out: only a well-designed **experiment** can show **causation**. An observational study can find a link (correlation) but cannot prove that one thing causes another.

Practice

Try these in order. The methods are all above.

2.1 State the difference between an observational study and an experiment. [2]

2.2 State the roles of the treatment group and the control group. [2]

2.3 State why subjects are randomly assigned to groups. [2]

2.4 State which kind of study can show that one thing causes another. [1]

2.5 Explain why a control group is needed to judge a new medicine. [2]

3. Alzheimer's disease results in a loss of cognitive ability beyond what is expected with typical aging. A local newspaper published an article with the following headline.

Study Finds Strong Association Between Smoking and Alzheimer's

The article reported that a study tracked the medical histories of 21,123 men and women for 23 years. The article stated that, for those who smoked at least two packs of cigarettes a day, the risk of developing Alzheimer's disease was 2.57 times the risk for those who did not smoke.

(a) Identify the explanatory and response variables in the study.

Explanatory variable:

Response variable:

(b) Is the study described in the article an observational study or an experiment? Explain.

(c) Exercise status (regular weekly exercise versus no regular weekly exercise) was mentioned in the article as a possible confounding variable. Explain how exercise status could be a confounding variable in the study.

3. Alzheimer's disease results in a loss of cognitive ability beyond what is expected with typical aging. A local newspaper published an article with the following headline.

Study Finds Strong Association Between Smoking and Alzheimer's

The article reported that a study tracked the medical histories of 21,123 men and women for 23 years. The article stated that, for those who smoked at least two packs of cigarettes a day, the risk of developing Alzheimer's disease was 2.57 times the risk for those who did not smoke.

(a) Identify the explanatory and response variables in the study.

Explanatory variable:

Response variable:

(b) Is the study described in the article an observational study or an experiment? Explain.

(c) Exercise status (regular weekly exercise versus no regular weekly exercise) was mentioned in the article as a possible confounding variable. Explain how exercise status could be a confounding variable in the study.

2. Researchers are investigating the effectiveness of using a fungus to control the spread of an insect that destroys trees. The researchers will create four different concentrations of fungus mixtures: 0 milliliters per liter (ml/L), 1.25 ml/L, 2.5 ml/L, and 3.75 ml/L. An equal number of the insects will be placed into 20 individual containers. The group of insects in each container will be sprayed with one of the four mixtures, and the researchers will record the number of insects that are still alive in each container one week after spraying.

(a) Identify the treatments, experimental units, and response variable of the experiment.

Treatments:

Experimental units:

Response variable:

(b) Does the experiment have a control group? Explain your answer.

(c) Describe how the treatments can be randomly assigned to the experimental units so that each treatment has the same number of units.

4. The anterior cruciate ligament (ACL) is one of the ligaments that help stabilize the knee. Surgery is often recommended if the ACL is completely torn, and recovery time from the surgery can be lengthy. A medical center developed a new surgical procedure designed to reduce the average recovery time from the surgery. To test the effectiveness of the new procedure, a study was conducted in which 210 patients needing surgery to repair a torn ACL were randomly assigned to receive either the standard procedure or the new procedure.

   (a) Based on the design of the study, would a statistically significant result allow the medical center to conclude that the new procedure causes a reduction in recovery time compared to the standard procedure, for patients similar to those in the study? Explain your answer.

   (b) Summary statistics on the recovery times from the surgery are shown in the table.

Type of Procedure	Sample Size	Mean Recovery Time (days)	Standard Deviation Recovery Time (days)
Standard	110	217	34
New	100	186	29

   Do the data provide convincing statistical evidence that those who receive the new procedure will have less recovery time from the surgery, on average, than those who receive the standard procedure, for patients similar to those in the study?

AP STATISTICS • HOLIDAY HOMEWORK 假期作业

4 Basic probability – Part 1

Name: ____________ Class: ________ Date: __________

Vocabulary 词汇

Write each word three times. 把每个单词抄写三遍。

English	中文	Write it 3 times (English)
Probability	概率	

Recall

You have a sense of chance already:

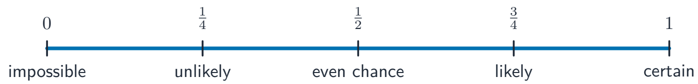
- A coin lands heads about half the time.
- A rolled die is equally likely to show 1 to 6.
- Some events are more likely than others.

This sheet lines up the basics of probability. Each idea has a worked example.

New ideas

1 The probability scale

Probability 概率 measures how likely an event is, from 0 (impossible) to 1 (certain).



2 Equally likely outcomes

When outcomes are equally likely,

$$P(\text{event}) = \frac{\text{favourable outcomes}}{\text{total outcomes}}.$$

Worked example. Rolling a 4 on a die: $P = \frac{1}{6}$. Rolling an even number: $P = \frac{3}{6} = \frac{1}{2}$.

3 Complement

The chance an event does **not** happen is 1 minus the chance it does: $P(\text{not } A) = 1 - P(A)$.

Worked example. If $P(\text{rain}) = 0.3$, then $P(\text{no rain}) = 1 - 0.3 = 0.7$.

4 Combining events

For two independent events (one does not affect the other), the chance **both** happen is the product of their chances.

Watch out: a probability is always between 0 and 1 (or 0% and 100%). An answer above 1 or below 0 is a sure sign of a mistake.

Practice

Try these in order. The methods are all above.

2.1 State what a probability of 0 and of 1 each mean. [2]

2.2 A die is rolled. Find the probability of rolling a 5. [2]

2.3 Find the probability of rolling an odd number on a die. [2]

2.4 If $P(\text{win}) = 0.4$, find $P(\text{not win})$. [2]

2.5 State the range that any probability must lie in. [1]

2. Nine sales representatives, 6 men and 3 women, at a small company wanted to attend a national convention. There were only enough travel funds to send 3 people. The manager selected 3 people to attend and stated that the people were selected at random. The 3 people selected were women. There were concerns that no men were selected to attend the convention.

(a) Calculate the probability that randomly selecting 3 people from a group of 6 men and 3 women will result in selecting 3 women.

(b) Based on your answer to part (a), is there reason to doubt the manager's claim that the 3 people were selected at random? Explain.

(c) An alternative to calculating the exact probability is to conduct a simulation to estimate the probability. A proposed simulation process is described below.

> Each trial in the simulation consists of rolling three fair, six-sided dice, one die for each of the convention attendees. For each die, rolling a 1, 2, 3, or 4 represents selecting a man; rolling a 5 or 6 represents selecting a woman. After 1,000 trials, the number of times the dice indicate selecting 3 women is recorded.

Does the proposed process correctly simulate the random selection of 3 women from a group of 9 people consisting of 6 men and 3 women? Explain why or why not.

2. Nine sales representatives, 6 men and 3 women, at a small company wanted to attend a national convention. There were only enough travel funds to send 3 people. The manager selected 3 people to attend and stated that the people were selected at random. The 3 people selected were women. There were concerns that no men were selected to attend the convention.

   (a) Calculate the probability that randomly selecting 3 people from a group of 6 men and 3 women will result in selecting 3 women.

   (b) Based on your answer to part (a), is there reason to doubt the manager's claim that the 3 people were selected at random? Explain.

   (c) An alternative to calculating the exact probability is to conduct a simulation to estimate the probability. A proposed simulation process is described below.

   > Each trial in the simulation consists of rolling three fair, six-sided dice, one die for each of the convention attendees. For each die, rolling a 1, 2, 3, or 4 represents selecting a man; rolling a 5 or 6 represents selecting a woman. After 1,000 trials, the number of times the dice indicate selecting 3 women is recorded.

   Does the proposed process correctly simulate the random selection of 3 women from a group of 9 people consisting of 6 men and 3 women? Explain why or why not.

Name: AP Statistics: **2014**

2. Nine sales representatives, 6 men and 3 women, at a small company wanted to attend a national convention. There were only enough travel funds to send 3 people. The manager selected 3 people to attend and stated that the people were selected at random. The 3 people selected were women. There were concerns that no men were selected to attend the convention.

   (a) Calculate the probability that randomly selecting 3 people from a group of 6 men and 3 women will result in selecting 3 women.

   (b) Based on your answer to part (a), is there reason to doubt the manager's claim that the 3 people were selected at random? Explain.

   (c) An alternative to calculating the exact probability is to conduct a simulation to estimate the probability. A proposed simulation process is described below.

   > Each trial in the simulation consists of rolling three fair, six-sided dice, one die for each of the convention attendees. For each die, rolling a 1, 2, 3, or 4 represents selecting a man; rolling a 5 or 6 represents selecting a woman. After 1,000 trials, the number of times the dice indicate selecting 3 women is recorded.

   Does the proposed process correctly simulate the random selection of 3 women from a group of 9 people consisting of 6 men and 3 women? Explain why or why not.

5. A researcher is investigating whether there is an association among professional athletes between age-group (in years) and type of sport played (basketball, football, baseball). The age-group and type of sport played for all 4,193 professional athletes in these sports for a recent year are given in the two-way table.

Age-Group by Type of Sport Played

Age (years)	Basketball	Football	Baseball	Total
Age < 25	232	807	259	1,298
25 ≤ Age < 30	175	1,326	620	2,121
30 ≤ Age < 35	90	287	276	653
35 ≤ Age	19	41	61	121
Total	516	2,461	1,216	4,193

Part A

Consider the two-way table.

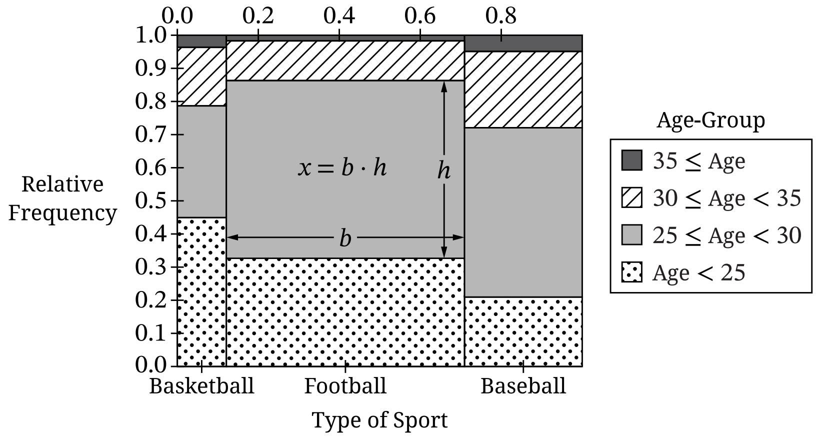
i. What is the probability a randomly selected professional athlete is a football player? Show your work.

ii. What is the probability a randomly selected professional athlete is in the age-group of "25 ≤ Age < 30" given they are a football player? Show your work.

Part B

A mosaic plot was constructed using the information in the two-way table. Use the mosaic plot to answer the following questions.

Relative Frequency of Age-Group by Type of Sport Played



i. The b and h displayed in the mosaic plot each represent a probability calculated in part A. Which probability does b correspond to: the probability calculated in part A (i) or the probability calculated in part A (ii)?

ii. What probability does the x displayed in the mosaic plot represent in context?

Part C

Use the information to answer the following questions.

i. Are the events "Baseball" and "35 ≤ Age" mutually exclusive events? Explain.

ii. Are the events "Baseball" and "35 ≤ Age" independent events? Show your work.

Part D

Consider the data of all professional athletes from the two-way table. Determine if it is appropriate to carry out a chi-square test for independence to investigate whether there is an association between age-group and sport played using these data. Explain your answer.

AP Statistics: **2018**

3. Approximately 3.5 percent of all children born in a certain region are from multiple births (that is, twins, triplets, etc.). Of the children born in the region who are from multiple births, 22 percent are left-handed. Of the children born in the region who are from single births, 11 percent are left-handed.

(a) What is the probability that a randomly selected child born in the region is left-handed?

(b) What is the probability that a randomly selected child born in the region is a child from a multiple birth, given that the child selected is left-handed?

(c) A random sample of 20 children born in the region will be selected. What is the probability that the sample will have at least 3 children who are left-handed?

AP Statistics: **2017**

3. A grocery store purchases melons from two distributors, J and K. Distributor J provides melons from organic farms. The distribution of the diameters of the melons from Distributor J is approximately normal with mean 133 millimeters (mm) and standard deviation 5 mm.

(a) For a melon selected at random from Distributor J, what is the probability that the melon will have a diameter greater than 137 mm?

Distributor K provides melons from nonorganic farms. The probability is 0.8413 that a melon selected at random from Distributor K will have a diameter greater than 137 mm. For all the melons at the grocery store, 70 percent of the melons are provided by Distributor J and 30 percent are provided by Distributor K.

(b) For a melon selected at random from the grocery store, what is the probability that the melon will have a diameter greater than 137 mm?

(c) Given that a melon selected at random from the grocery store has a diameter greater than 137 mm, what is the probability that the melon will be from Distributor J?

AP Statistics: **2017**

3. A grocery store purchases melons from two distributors, J and K. Distributor J provides melons from organic farms. The distribution of the diameters of the melons from Distributor J is approximately normal with mean 133 millimeters (mm) and standard deviation 5 mm.

(a) For a melon selected at random from Distributor J, what is the probability that the melon will have a diameter greater than 137 mm?

Distributor K provides melons from nonorganic farms. The probability is 0.8413 that a melon selected at random from Distributor K will have a diameter greater than 137 mm. For all the melons at the grocery store, 70 percent of the melons are provided by Distributor J and 30 percent are provided by Distributor K.

(b) For a melon selected at random from the grocery store, what is the probability that the melon will have a diameter greater than 137 mm?

(c) Given that a melon selected at random from the grocery store has a diameter greater than 137 mm, what is the probability that the melon will be from Distributor J?

AP Statistics: **2016**

4. A company manufactures model rockets that require igniters to launch. Once an igniter is used to launch a rocket, the igniter cannot be reused. Sometimes an igniter fails to operate correctly, and the rocket does not launch. The company estimates that the overall failure rate, defined as the percent of all igniters that fail to operate correctly, is 15 percent.

A company engineer develops a new igniter, called the super igniter, with the intent of lowering the failure rate. To test the performance of the super igniters, the engineer uses the following process.

Step 1: One super igniter is selected at random and used in a rocket.

Step 2: If the rocket launches, another super igniter is selected at random and used in a rocket.

Step 2 is repeated until the process stops. The process stops when a super igniter fails to operate correctly or 32 super igniters have successfully launched rockets, whichever comes first. Assume that super igniter failures are independent.

(a) If the failure rate of the super igniters is 15 percent, what is the probability that the first 30 super igniters selected using the testing process successfully launch rockets?

(b) Given that the first 30 super igniters successfully launch rockets, what is the probability that the first failure occurs on the thirty-first or the thirty-second super igniter tested if the failure rate of the super igniters is 15 percent?

(c) Given that the first 30 super igniters successfully launch rockets, is it reasonable to believe that the failure rate of the super igniters is less than 15 percent? Explain.

4 Random variables and distributions – Part 2

Name: __________ Class: _______ Date: ________

Recall

In Part A you found probabilities of single events. Now handle chance that varies:

- A number that depends on chance (like a dice total) is a random variable.
- Its possible values each have a probability.

Each idea has a worked example before the Practice.

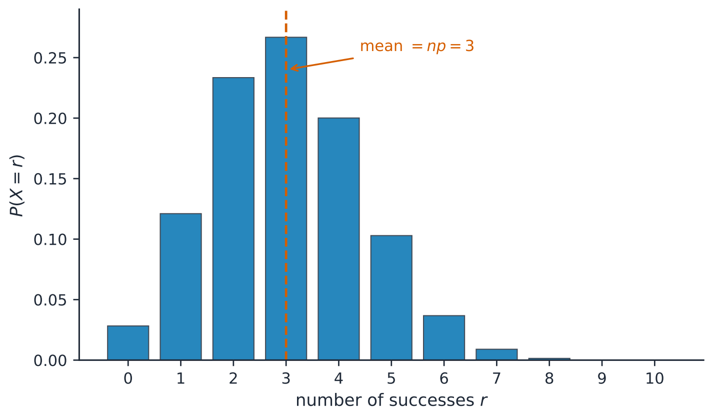
New ideas

1 Random variables

A **random variable** is a number whose value depends on chance – like the number of heads in 3 coin flips.

2 Probability distributions

A **probability distribution** lists each possible value and its probability. All the probabilities add up to 1.



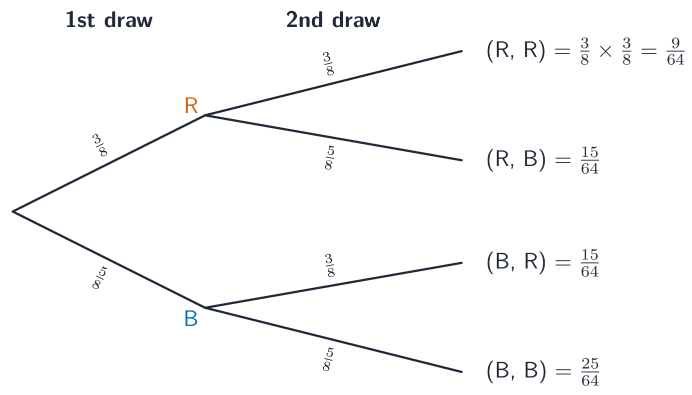
3 Expected value

The **expected value** (mean) is the long-run average: multiply each value by its probability and add.

Worked example. A game pays \$0 with probability 0.5 and \$10 with probability 0.5: expected value $= 0(0.5) + 10(0.5) = \$5$.

4 Tree diagrams

A **tree diagram** maps out the outcomes of several stages, so you can multiply probabilities along the branches.



Watch out: in any probability distribution, all the probabilities must add up to exactly 1. If your listed probabilities do not sum to 1, something is missing or wrong.

Practice

Try these in order. The methods are all above.

2.1 State what a random variable is. [2]

2.2 State what all the probabilities in a distribution must add up to. [1]

2.3 A prize is \$0 with probability 0.8 and \$20 with probability 0.2. Find the expected value. [3]

2.4 A distribution lists probabilities 0.3, 0.3, and 0.3. State why something must be wrong. [2]

2.5 State what a tree diagram is used for. [1]

3. A shopping mall has three automated teller machines (ATMs). Because the machines receive heavy use, they sometimes stop working and need to be repaired. Let the random variable X represent the number of ATMs that are working when the mall opens on a randomly selected day. The table shows the probability distribution of X.

Number of ATMs working when the mall opens	0	1	2	3
Probability	0.15	0.21	0.40	0.24

(a) What is the probability that at least one ATM is working when the mall opens?

(b) What is the expected value of the number of ATMs that are working when the mall opens?

(c) What is the probability that all three ATMs are working when the mall opens, given that at least one ATM is working?

(d) Given that at least one ATM is working when the mall opens, would the expected value of the number of ATMs that are working be less than, equal to, or greater than the expected value from part (b) ? Explain.

3. A shopping mall has three automated teller machines (ATMs). Because the machines receive heavy use, they sometimes stop working and need to be repaired. Let the random variable X represent the number of ATMs that are working when the mall opens on a randomly selected day. The table shows the probability distribution of X.

Number of ATMs working when the mall opens	0	1	2	3
Probability	0.15	0.21	0.40	0.24

(a) What is the probability that at least one ATM is working when the mall opens?

(b) What is the expected value of the number of ATMs that are working when the mall opens?

(c) What is the probability that all three ATMs are working when the mall opens, given that at least one ATM is working?

(d) Given that at least one ATM is working when the mall opens, would the expected value of the number of ATMs that are working be less than, equal to, or greater than the expected value from part (b) ? Explain.

Descriptive Statistics

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{\sum x_i}{n} \qquad s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$\hat{y} = a + bx \qquad \bar{y} = a + b\bar{x}$$

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \qquad b = r\frac{s_y}{s_x}$$

Probability and Distributions

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Probability Distribution	Mean	Standard Deviation
Discrete random variable, X	$\mu_X = E(X) = \sum x_i P(x_i)$	$\sigma_X = \sqrt{\sum (x_i - \mu_X)^2 P(x_i)}$
If X has a **binomial** distribution with parameters n and p, then: $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ where $x = 0, 1, 2, 3, \ldots, n$	$\mu_X = np$	$\sigma_X = \sqrt{np(1-p)}$
If X has a **geometric** distribution with parameter p, then: $P(X = x) = (1-p)^{x-1} p$ where $x = 1, 2, 3, \ldots$	$\mu_X = \frac{1}{p}$	$\sigma_X = \frac{\sqrt{1-p}}{p}$

.

Sampling Distributions and Inferential Statistics

Standardized test statistic: $\dfrac{\text{statistic} - \text{parameter}}{\text{standard error of the statistic}}$

Confidence interval: statistic $\pm$ (critical value)(standard error of statistic)

Chi-square statistic: $\chi^2 = \sum \dfrac{(\text{observed} - \text{expected})^2}{\text{expected}}$

Sampling distributions for proportions:

Random Variable	Parameters of Sampling Distribution		Standard Error* of Sample Statistic
For one population: $\hat{p}$	$\mu_{\hat{p}} = p$	$\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$	$s_{\hat{p}} = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$
For two populations: $\hat{p}_1 - \hat{p}_2$	$\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$	$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$	$s_{\hat{p}_1-\hat{p}_2} = \sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ When $p_1 = p_2$ is assumed: $s_{\hat{p}_1-\hat{p}_2} = \sqrt{\hat{p}_c(1-\hat{p}_c)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}$ where $\hat{p}_c = \dfrac{X_1 + X_2}{n_1 + n_2}$

Sampling distributions for means:

Random Variable	Parameters of Sampling Distribution		Standard Error* of Sample Statistic
For one population: $\overline{X}$	$\mu_{\overline{X}} = \mu$	$\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$	$s_{\overline{X}} = \dfrac{s}{\sqrt{n}}$
For two populations: $\overline{X}_1 - \overline{X}_2$	$\mu_{\overline{X}_1-\overline{X}_2} = \mu_1 - \mu_2$	$\sigma_{\overline{X}_1-\overline{X}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$	$s_{\overline{X}_1-\overline{X}_2} = \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$

Sampling distributions for simple linear regression:

Random Variable	Parameters of Sampling Distribution		Standard Error* of Sample Statistic
For slope: b	$\mu_b = \beta$	$\sigma_b = \dfrac{\sigma}{\sigma_x\sqrt{n}}$, where $\sigma_x = \sqrt{\dfrac{\Sigma(x_i - \bar{x})^2}{n}}$	$s_b = \dfrac{s}{s_x\sqrt{n-1}}$, where $s = \sqrt{\dfrac{\Sigma(y_i - \hat{y}_i)^2}{n-2}}$ and $s_x = \sqrt{\dfrac{\Sigma(x_i - \bar{x})^2}{n-1}}$

andard deviation is a measurement of variability from the theoretical population. Standard error is the estimate of the standard deviation. If the ndard deviation of the statistic is assumed to be known, then the standard deviation should be used instead of the standard error.

3. Ms. Fey is a manager at a restaurant. To improve the dining experience for her customers, she uses a digital music service to create a playlist of songs that will be played in the restaurant. The playlist contains 1,000 songs and consists of four different types of music in the following quantities: 200 country songs, 400 pop songs, 100 rock songs, and 300 jazz songs. The digital music service will select songs at random from the playlist to be played in the restaurant. Any song can be replayed at any time.

A.

i. Suppose one song is selected at random to be played. What is the probability that the song is a rock song? Show your work.

ii. Suppose two songs are selected at random to be played. What is the probability that both songs are rock songs? Show your work.

B. In every one-hour period, 20 songs will be played at random and any song can be replayed at any time. Ms. Fey is interested in how many rock songs will be played in a typical one-hour period.

i. Define the random variable of interest to Ms. Fey, and state how the random variable is distributed.

ii. What is the expected value for the random variable in part B (i)? Show your work.

C. Recall that in every one-hour period, 20 songs will be played at random and any song can be replayed at any time.

i. Determine the probability that 4 or more rock songs in a particular one-hour period will be played. Show your work.

ii. Suppose 4 rock songs are played during a particular one-hour period. Does this provide strong evidence that the song selection process was not truly random? Justify your answer without performing an inference procedure.

3. Approximately 3.5 percent of all children born in a certain region are from multiple births (that is, twins, triplets, etc.). Of the children born in the region who are from multiple births, 22 percent are left-handed. Of the children born in the region who are from single births, 11 percent are left-handed.

(a) What is the probability that a randomly selected child born in the region is left-handed?

(b) What is the probability that a randomly selected child born in the region is a child from a multiple birth, given that the child selected is left-handed?

(c) A random sample of 20 children born in the region will be selected. What is the probability that the sample will have at least 3 children who are left-handed?

3. Approximately 3.5 percent of all children born in a certain region are from multiple births (that is, twins, triplets, etc.). Of the children born in the region who are from multiple births, 22 percent are left-handed. Of the children born in the region who are from single births, 11 percent are left-handed.

   (a) What is the probability that a randomly selected child born in the region is left-handed?

   (b) What is the probability that a randomly selected child born in the region is a child from a multiple birth, given that the child selected is left-handed?

   (c) A random sample of 20 children born in the region will be selected. What is the probability that the sample will have at least 3 children who are left-handed?

AP Statistics: **2016**

4. A company manufactures model rockets that require igniters to launch. Once an igniter is used to launch a rocket, the igniter cannot be reused. Sometimes an igniter fails to operate correctly, and the rocket does not launch. The company estimates that the overall failure rate, defined as the percent of all igniters that fail to operate correctly, is 15 percent.

   A company engineer develops a new igniter, called the super igniter, with the intent of lowering the failure rate. To test the performance of the super igniters, the engineer uses the following process.

   Step 1: One super igniter is selected at random and used in a rocket.

   Step 2: If the rocket launches, another super igniter is selected at random and used in a rocket.

   Step 2 is repeated until the process stops. The process stops when a super igniter fails to operate correctly or 32 super igniters have successfully launched rockets, whichever comes first. Assume that super igniter failures are independent.

   (a) If the failure rate of the super igniters is 15 percent, what is the probability that the first 30 super igniters selected using the testing process successfully launch rockets?

   (b) Given that the first 30 super igniters successfully launch rockets, what is the probability that the first failure occurs on the thirty-first or the thirty-second super igniter tested if the failure rate of the super igniters is 15 percent?

   (c) Given that the first 30 super igniters successfully launch rockets, is it reasonable to believe that the failure rate of the super igniters is less than 15 percent? Explain.

4. A company manufactures model rockets that require igniters to launch. Once an igniter is used to launch a rocket, the igniter cannot be reused. Sometimes an igniter fails to operate correctly, and the rocket does not launch. The company estimates that the overall failure rate, defined as the percent of all igniters that fail to operate correctly, is 15 percent.

   A company engineer develops a new igniter, called the super igniter, with the intent of lowering the failure rate. To test the performance of the super igniters, the engineer uses the following process.

   Step 1: One super igniter is selected at random and used in a rocket.

   Step 2: If the rocket launches, another super igniter is selected at random and used in a rocket.

   Step 2 is repeated until the process stops. The process stops when a super igniter fails to operate correctly or 32 super igniters have successfully launched rockets, whichever comes first. Assume that super igniter failures are independent.

   (a) If the failure rate of the super igniters is 15 percent, what is the probability that the first 30 super igniters selected using the testing process successfully launch rockets?

   (b) Given that the first 30 super igniters successfully launch rockets, what is the probability that the first failure occurs on the thirty-first or the thirty-second super igniter tested if the failure rate of the super igniters is 15 percent?

   (c) Given that the first 30 super igniters successfully launch rockets, is it reasonable to believe that the failure rate of the super igniters is less than 15 percent? Explain.

5 Sampling variability – Part 1

Name: ____________ Class: ________ Date: __________

Vocabulary 词汇

Write each word three times. 把每个单词抄写三遍。

English	中文	Write it 3 times (English)
sampling distribution	抽样分布	______ ______ ______

Recall

You know a sample is used to estimate a whole population:

- But two different samples will give slightly different answers.
- The sample mean is not exactly the true mean – it varies.

This sheet lines up how sample results vary. Each idea has a worked example.

New ideas

■ 1 Statistics vary

A **statistic** (like a sample mean) changes from one sample to the next, just by chance. This is **sampling variability**.

■ 2 The sampling distribution

The **sampling distribution** 抽样分布 is the distribution of a statistic over many, many samples – it shows how much the statistic bounces around.

■ 3 Centered on the truth

For a fair sample, the sampling distribution is **centered** on the true population value – so on average, the sample gives the right answer.

■ 4 Bigger samples, less spread

Worked example. Estimating an average with a sample of 100 gives a much less variable result than a sample of 10 – larger samples give more reliable estimates.

Watch out: the sample mean is almost never **exactly** equal to the population mean – it varies by chance. What we can say is how **close** it is likely to be, and larger

samples make it closer.

Practice

Try these in order. The methods are all above.

2.1 State what "sampling variability" means. [2]

2.2 State what a sampling distribution shows. [2]

2.3 State where the sampling distribution of a fair sample is centered. [1]

2.4 State the effect of a larger sample on the variability of an estimate. [2]

2.5 Explain why two different samples from the same population can give different means. [2]

STATISTICS
SECTION II
Part B
Question 6

Spend about 25 minutes on this part of the exam.

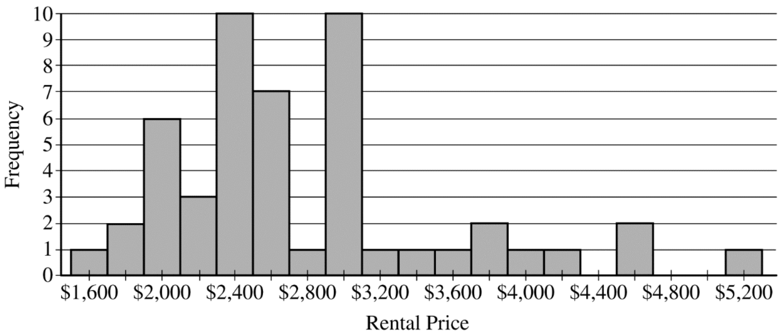
Percent of Section II score—25

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

6. Emma is moving to a large city and is investigating typical monthly rental prices of available one-bedroom apartments. She obtained a random sample of rental prices for 50 one-bedroom apartments taken from a Web site where people voluntarily list available apartments.

   (a) Describe the population for which it is appropriate for Emma to generalize the results from her sample.

   The distribution of the 50 rental prices of the available apartments is shown in the following histogram.



   (b) Emma wants to estimate the typical rental price of a one-bedroom apartment in the city. Based on the distribution shown, what is a disadvantage of using the mean rather than the median as an estimate of the typical rental price?

   (c) Instead of using the sample median as the point estimate for the population median, Emma wants to use an interval estimate. However, computing an interval estimate requires knowing the sampling distribution of the sample median for samples of size 50. Emma has one point, her sample median, in that sampling distribution. Using information about rental prices that are available on the Web site, describe how someone could develop a theoretical sampling distribution of the sample median for samples of size 50.

   Because Emma does not have the resources to develop the theoretical sampling distribution, she estimates the sampling distribution of the sample median using a process called bootstrapping. In the bootstrapping process, a computer program performs the following steps.

   - Take a random sample, with replacement, of size 50 from the original sample.
   - Calculate and record the median of the sample.
   - Repeat the process to obtain a total of 15,000 medians.

Emma ran the bootstrap process, and the following frequency table is the bootstrap distribution showing her results of generating 15,000 medians.

Bootstrap Distribution of Medians					
Median	Frequency	Median	Frequency	Median	Frequency
2,345	1	2,585	1	2,825	247
2,390	13	2,587.5	171	2,837.5	7
2,395	18	2,600	22	2,847.5	1
2,400	56	2,612.5	1,190	2,872.5	317
2,445	4	2,625	174	2,885	10
2,447.5	56	2,672.5	5	2,950	700
2,450	55	2,675	1,924	2,962.5	93
2,475	3	2,687.5	1,341	2,972.5	6
2,495	66	2,700	2,825	2,975	65
2,497.5	136	2,735	35	2,985	12
2,500	1,899	2,747.5	619	2,987.5	1
2,522.5	2	2,750	2	2,995	6
2,525	945	2,795	278	3,000	2
2,550	1,673	2,812.5	16	3,062.5	3

The bootstrap distribution provides an approximation of the sampling distribution of the sample median. A confidence interval for the median can be constructed using a percentage of the values in the middle of the bootstrap distribution.

(d) Use the frequency table to find the following.

   (i) Value of the 5th percentile:

   (ii) Value of the 95th percentile:

(e) Find the percentage of bootstrap medians in the table that are equal to or between the values found in part (d).

(f) Use your values from parts (d) and (e) to construct and interpret a confidence interval for the median rental price.

STOP

END OF EXAM

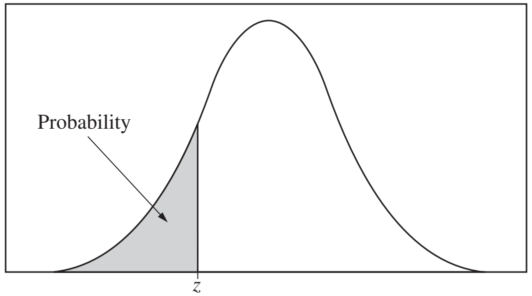


Table entry for z is the probability lying below z.

Table A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

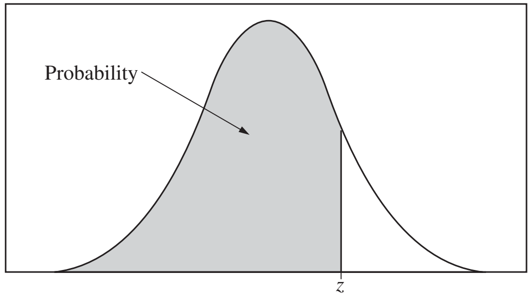


Table entry for z is the probability lying below z.

Table A (*Continued*)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Table entry for p and C is the point t^* with probability p lying above it and probability C lying between $-t^*$ and t^*.

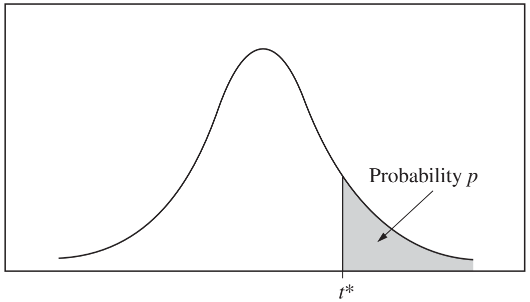


Table B ***t* distribution critical values**

df	Tail probability p: .25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
Confidence level C	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%

Table entry for p is the point (χ^2) with probability p lying above it.

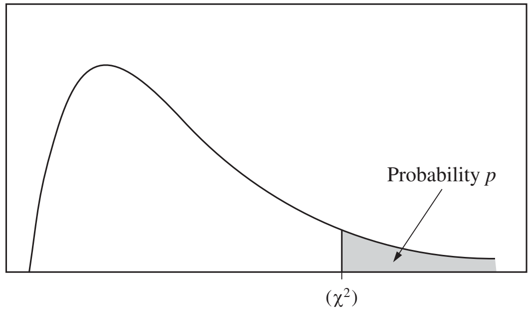


Table C χ^2 **critical values**

df	Tail probability p: .25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.7
80	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8	128.3
100	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4	153.2

2. An environmental science teacher at a high school with a large population of students wanted to estimate the proportion of students at the school who regularly recycle plastic bottles. The teacher selected a random sample of students at the school to survey. Each selected student went into the teacher's office, one at a time, and was asked to respond yes or no to the following question.

Do you regularly recycle plastic bottles?

Based on the responses, a 95 percent confidence interval for the proportion of all students at the school who would respond yes to the question was calculated as $(0.584, 0.816)$.

(a) How many students were in the sample selected by the environmental science teacher?

(b) Given the method used by the environmental science teacher to collect the responses, explain how bias might have been introduced and describe how the bias might affect the point estimate of the proportion of all students at the school who would respond yes to the question.

(c) The statistics teacher at the high school was concerned about the potential bias in the survey. To obtain a potentially less biased estimate of the proportion, the statistics teacher used an alternate method for collecting student responses. A random sample of 300 students was selected, and each student was given the following instructions on how to respond to the question.

- In private, flip a fair coin.
- If heads, you must respond no, regardless of whether you regularly recycle.
- If tails, please truthfully respond yes or no.

(i) What is the expected number of students from the sample of 300 who would be required to respond no because the coin flip resulted in heads?

(ii) The results of the sample showed that 213 of the 300 selected students responded no. Based on the results of the sample, give a point estimate for the proportion of all students at the high school who would respond yes to the question.

5 The central limit theorem – Part 2

Name: ____________ Class: ________ Date: __________

Vocabulary 词汇

Write each word three times. 把每个单词抄写三遍。

English	中文	Write it 3 times (English)
central limit theorem	中心极限定理	____________ ____________ ____________

Recall

In Part A you saw that sample means vary. There is a beautiful pattern to how they vary:

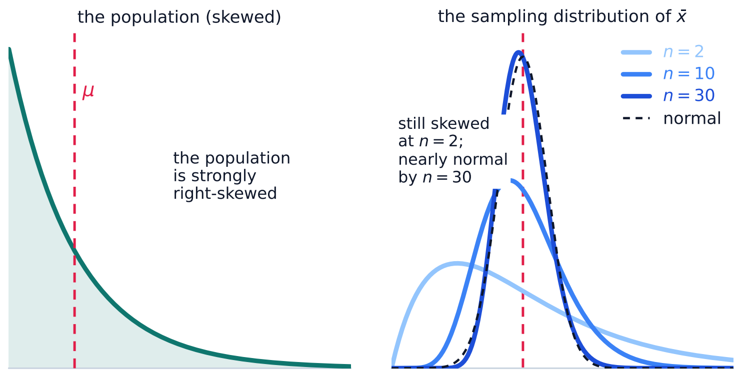
- Even if the original data are not bell-shaped, the sample means often are.
- This pattern is the key that makes statistics work.

Each idea has a worked example before the Practice.

New ideas

1 The central limit theorem

The **central limit theorem** 中心极限定理 says: for a large enough sample, the distribution of the **sample mean** is approximately **normal** (bell-shaped), centered on the population mean.



CLT: for large n, $\bar{x} \approx \text{Normal}(\mu,\ \sigma/\sqrt{n})$ even when the population is not normal

■ 2 Even from non-normal data

Remarkably, this holds **even if the original population is not normal** – the averaging smooths things into a bell shape.

■ 3 Larger samples, tighter bell

The bigger the sample, the narrower and more normal the distribution of the sample mean becomes.

■ 4 Why it matters

Because the sample mean is approximately normal, we can use the familiar bell curve (and the 68–95–99.7 rule) to say how confident we are in an estimate – the basis of all the inference that follows.

Watch out: the central limit theorem is about the distribution of the **sample mean**, not the individual data. The raw data can be skewed while the sample means are still bell-shaped.

Practice

Try these in order. The methods are all above.

2.1 State what the central limit theorem says about the sample mean. [2]

2.2 State the surprising part – what it does not require of the population. [2]

2.3 State what happens to the sample mean's distribution as the sample gets larger. [2]

2.4 State why the central limit theorem is so useful in statistics. [2]

2.5 State whether the theorem is about the individual data or the sample mean. [1]

2. Nine sales representatives, 6 men and 3 women, at a small company wanted to attend a national convention. There were only enough travel funds to send 3 people. The manager selected 3 people to attend and stated that the people were selected at random. The 3 people selected were women. There were concerns that no men were selected to attend the convention.

(a) Calculate the probability that randomly selecting 3 people from a group of 6 men and 3 women will result in selecting 3 women.

(b) Based on your answer to part (a), is there reason to doubt the manager's claim that the 3 people were selected at random? Explain.

(c) An alternative to calculating the exact probability is to conduct a simulation to estimate the probability. A proposed simulation process is described below.

Each trial in the simulation consists of rolling three fair, six-sided dice, one die for each of the convention attendees. For each die, rolling a 1, 2, 3, or 4 represents selecting a man; rolling a 5 or 6 represents selecting a woman. After 1,000 trials, the number of times the dice indicate selecting 3 women is recorded.

Does the proposed process correctly simulate the random selection of 3 women from a group of 9 people consisting of 6 men and 3 women? Explain why or why not.

---

3. Schools in a certain state receive funding based on the number of students who attend the school. To determine the number of students who attend a school, one school day is selected at random and the number of students in attendance that day is counted and used for funding purposes. The daily number of absences at High School A in the state is approximately normally distributed with mean of 120 students and standard deviation of 10.5 students.

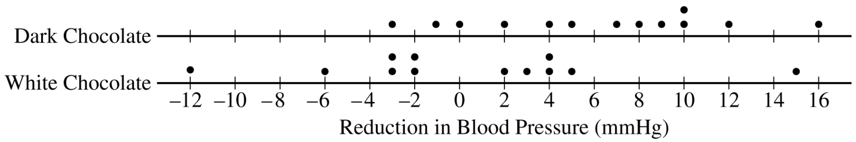
(a) If more than 140 students are absent on the day the attendance count is taken for funding purposes, the school will lose some of its state funding in the subsequent year. Approximately what is the probability that High School A will lose some state funding?

(b) The principals' association in the state suggests that instead of choosing one day at random, the state should choose 3 days at random. With the suggested plan, High School A would lose some of its state funding in the subsequent year if the mean number of students absent for the 3 days is greater than 140. Would High School A be more likely, less likely, or equally likely to lose funding using the suggested plan compared to the plan described in part (a)? Justify your choice.

(c) A typical school week consists of the days Monday, Tuesday, Wednesday, Thursday, and Friday. The principal at High School A believes that the number of absences tends to be greater on Mondays and Fridays, and there is concern that the school will lose state funding if the attendance count occurs on a Monday or Friday. If one school day is chosen at random from each of 3 typical school weeks, what is the probability that none of the 3 days chosen is a Tuesday, Wednesday, or Thursday?

. Studies have shown that foods rich in compounds known as flavonoids help lower blood pressure. Researchers conducted a study to investigate whether there was a greater reduction in blood pressure for people who consumed dark chocolate, which contains flavonoids, than people who consumed white chocolate, which does not contain flavonoids. Twenty-five healthy adults agreed to participate in the study and add 3.5 ounces of chocolate to their daily diets. Of the 25 participants, 13 were randomly assigned to the dark chocolate group and the rest were assigned to the white chocolate group. All participants had their blood pressure recorded, in millimeters of mercury (mmHg), before adding chocolate to their daily diets and again 30 days after adding chocolate to their daily diets.

The reduction in blood pressure (before minus after) for each of the participants in the two groups is shown in the dotplots below.



(a) Determine and compare the medians of the reduction in blood pressure for the two groups.

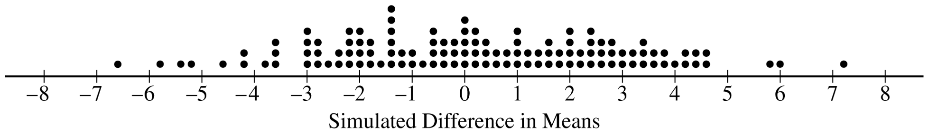
The researchers found the mean reduction in blood pressure for those who consumed dark chocolate is $\bar{x}_{dark} = 6.08$ mmHg and the mean reduction in blood pressure for those who consumed white chocolate is $\bar{x}_{white} = 0.42$ mmHg.

(b) One researcher indicated that because the difference in sample means of 5.66 mmHg is greater than 0 there is convincing statistical evidence to conclude that the population mean blood pressure reduction for those who consume dark chocolate is greater than for those who consume white chocolate. Why might the researcher's conclusion, based only on the difference in sample means of 5.66 mmHg, not necessarily be true?

GO ON TO THE NEXT PAGE.

Continue your response to **QUESTION 5** on this page.

A simulation was conducted to investigate whether there is a greater reduction of blood pressure for those who consume dark chocolate than for those who consume white chocolate. The simulation was conducted under the assumption that no difference exists. The results of 120 trials of the simulation are shown in the following dotplot.



(c) Use the results of the simulation to determine whether the results from the 25 participants in the study provide convincing statistical evidence, at a 5 percent level of significance, that adding dark chocolate to a daily diet will result in a greater reduction in blood pressure, on average, than adding white chocolate to a daily diet. Justify your answer.

GO ON TO THE NEXT PAGE.

6 Confidence intervals – Part 1

Name: ____________ Class: ________ Date: __________

Vocabulary 词汇

Write each word three times. 把每个单词抄写三遍。

English	中文	Write it 3 times (English)
confidence interval	置信区间	

Recall

You use a sample to estimate something about a population:

- A poll estimates the fraction of people who support an idea.
- But the estimate is not exact – how sure can we be?

This sheet introduces confidence intervals. Each idea has a worked example.

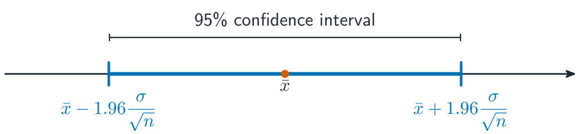
New ideas

■ 1 A range, not a single number

Because a sample estimate varies, we give a **range** likely to contain the true value, not just one number.

■ 2 The confidence interval

A **confidence interval** 置信区间 is an estimate plus-or-minus a margin of error: "the true value is probably in this range".



Worked example. A poll finds 60% support with a margin of error of ±3%, giving a confidence interval of 57% to 63%.

■ 3 Confidence level

A 95% **confidence level** means the method captures the true value 95% of the time it is used – we are 95% confident the range is right.

■ 4 Wider or narrower

A larger sample gives a **smaller** margin of error (a narrower, more precise interval); a higher confidence level gives a **wider** one.

Watch out: the confidence interval estimates the **population** value, not a single person. And "95% confident" refers to the method's long-run success rate, not a probability about this one interval.

Practice

Try these in order. The methods are all above.

2.1 State why we give a range instead of a single estimate. [2]

2.2 A poll finds 45% with a margin of error of ±4%. State the confidence interval. [2]

2.3 A poll finds 70% with a margin of error of ±2%. State the confidence interval. [2]

2.4 State what a larger sample does to the margin of error. [1]

2.5 State what a "95% confidence level" means. [2]

2. The manager of a local fast-food restaurant is concerned about customers who ask for a water cup when placing an order but fill the cup with a soft drink from the beverage fountain instead of filling the cup with water. The manager selected a random sample of 80 customers who asked for a water cup when placing an order and found that 23 of those customers filled the cup with a soft drink from the beverage fountain.

   (a) Construct and interpret a 95 percent confidence interval for the proportion of all customers who, having asked for a water cup when placing an order, will fill the cup with a soft drink from the beverage fountain.

   (b) The manager estimates that each customer who asks for a water cup but fills it with a soft drink costs the restaurant $0.25. Suppose that in the month of June 3,000 customers ask for a water cup when placing an order. Use the confidence interval constructed in part (a) to give an interval estimate for the cost to the restaurant for the month of June from the customers who ask for a water cup but fill the cup with a soft drink.

AP Statistics: **2016**

5. A polling agency showed the following two statements to a random sample of 1,048 adults in the United States.

   Environment statement: Protection of the environment should be given priority over economic growth.

   Economy statement: Economic growth should be given priority over protection of the environment.

   The order in which the statements were shown was randomly selected for each person in the sample. After reading the statements, each person was asked to choose the statement that was most consistent with his or her opinion. The results are shown in the table.

	Environment Statement	Economy Statement	No Preference
Percent of sample	58%	37%	5%

   (a) Assume the conditions for inference have been met. Construct and interpret a 95 percent confidence interval for the proportion of all adults in the United States who would have chosen the economy statement.

   (b) One of the conditions for inference that was met is that the number who chose the economy statement and the number who did not choose the economy statement are both greater than 10. Explain why it is necessary to satisfy that condition.

   (c) A suggestion was made to use a two-sample z-interval for a difference between proportions to investigate whether the difference in proportions between adults in the United States who would have chosen the environment statement and adults in the United States who would have chosen the economy statement is statistically significant. Is the two-sample z-interval for a difference between proportions an appropriate procedure to investigate the difference? Justify your answer.

2. The manager of a local fast-food restaurant is concerned about customers who ask for a water cup when placing an order but fill the cup with a soft drink from the beverage fountain instead of filling the cup with water. The manager selected a random sample of 80 customers who asked for a water cup when placing an order and found that 23 of those customers filled the cup with a soft drink from the beverage fountain.

(a) Construct and interpret a 95 percent confidence interval for the proportion of all customers who, having asked for a water cup when placing an order, will fill the cup with a soft drink from the beverage fountain.

(b) The manager estimates that each customer who asks for a water cup but fills it with a soft drink costs the restaurant \$0.25. Suppose that in the month of June 3,000 customers ask for a water cup when placing an order. Use the confidence interval constructed in part (a) to give an interval estimate for the cost to the restaurant for the month of June from the customers who ask for a water cup but fill the cup with a soft drink.

AP Statistics: **2015**

2. To increase business, the owner of a restaurant is running a promotion in which a customer's bill can be randomly selected to receive a discount. When a customer's bill is printed, a program in the cash register randomly determines whether the customer will receive a discount on the bill. The program was written to generate a discount with a probability of 0.2, that is, giving 20 percent of the bills a discount in the long run. However, the owner is concerned that the program has a mistake that results in the program not generating the intended long-run proportion of 0.2.

The owner selected a random sample of bills and found that only 15 percent of them received discounts. A confidence interval for p, the proportion of bills that will receive a discount in the long run, is 0.15 ± 0.06. All conditions for inference were met.

(a) Consider the confidence interval 0.15 ± 0.06.

(i) Does the confidence interval provide convincing statistical evidence that the program is not working as intended? Justify your answer.

(ii) Does the confidence interval provide convincing statistical evidence that the program generates the discount with a probability of 0.2 ? Justify your answer.

A second random sample of bills was taken that was four times the size of the original sample. In the second sample 15 percent of the bills received the discount.

(b) Determine the value of the margin of error based on the second sample of bills that would be used to compute an interval for p with the same confidence level as that of the original interval.

(c) Based on the margin of error in part (b) that was obtained from the second sample, what do you conclude about whether the program is working as intended? Justify your answer.

4. A software application (app) lets users enter questions to receive answers in the form of images, texts, or videos. Research indicates that 22 percent of high school students in Country W use the app to help them with their homework at least once per week. Karen is an AP Statistics student in Country W at a high school that has more than 2,000 students. She believes the proportion of all students at her school who use the app to help them with their homework at least once per week is greater than the proportion for her country. To investigate her belief, she took a simple random sample of 130 students from her school and found that 38 of the sampled students use the app to help them with their homework at least once per week.

Is there convincing statistical evidence, at a 0.05 significance level, to support Karen's belief? Justify your answer with the appropriate inference procedure.

4. A software application (app) lets users enter questions to receive answers in the form of images, texts, or videos. Research indicates that 22 percent of high school students in Country W use the app to help them with their homework at least once per week. Karen is an AP Statistics student in Country W at a high school that has more than 2,000 students. She believes the proportion of all students at her school who use the app to help them with their homework at least once per week is greater than the proportion for her country. To investigate her belief, she took a simple random sample of 130 students from her school and found that 38 of the sampled students use the app to help them with their homework at least once per week.

Is there convincing statistical evidence, at a 0.05 significance level, to support Karen's belief? Justify your answer with the appropriate inference procedure.

6 Hypothesis tests – Part 2

Name: ____________ Class: ________ Date: __________

Vocabulary 词汇

Write each word three times. 把每个单词抄写三遍。

English	中文	Write it 3 times (English)
hypothesis test	假设检验	____________ ____________ ____________

Recall

In Part A you estimated a value with a confidence interval. Now test a claim:

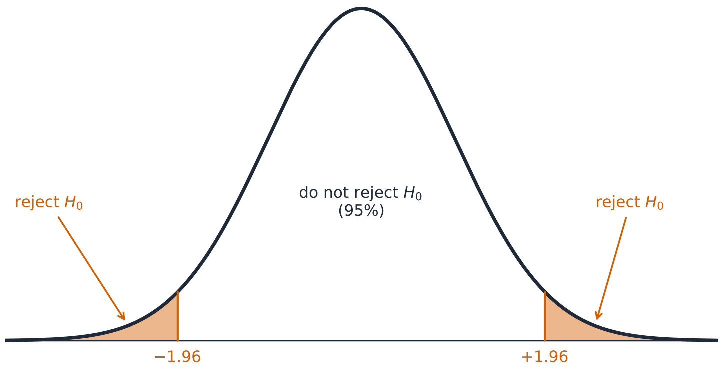
- Is a coin fair, or biased?
- Does the data give real evidence, or could it just be chance?

Each idea has a worked example before the Practice.

New ideas

1 A hypothesis test

A **hypothesis test** 假设检验 weighs evidence against a claim. It starts with a **null hypothesis** – "nothing unusual is going on" (for example, the coin is fair).



2 The p-value

The **p-value** is the chance of getting data at least this extreme **if the null hypothesis were true**. A **small** p-value means the data would be surprising, which is evidence against the null.

3 The decision

If the p-value is below a chosen level (often 0.05), we **reject** the null hypothesis – the effect is "statistically significant". Otherwise we do not reject it.

Worked example. A coin gives 9 heads in 10 flips. If the coin were fair, that is very unlikely (small p-value), so we have evidence it is biased.

4 What it does not prove

Failing to reject the null does **not** prove it true – it just means there was not enough evidence against it.

Watch out: a small p-value means the data are **surprising if the null is true** – it does not directly give "the probability the null is true". And "not significant" means "not enough evidence", not "proven no effect".

Practice

Try these in order. The methods are all above.

2.1 State what the null hypothesis usually represents. [2]

2.2 State what a p-value measures. [2]

2.3 State what a small p-value tells you. [2]

2.4 A test gives a p-value of 0.02 (below 0.05). State the decision. [1]

2.5 Explain why failing to reject the null does not prove it true. [2]

5. According to a 2017 national survey in Country B, the mean number of bedrooms in newly built houses was 2.9. Rodney, a researcher, believes the mean number of bedrooms in newly built houses in the country was different in 2024 than it was in 2017. To investigate his belief, he took a large random sample of newly built houses in Country B in 2024 and recorded the number of bedrooms in each house. The distribution of the number of bedrooms for the sampled houses is summarized in the table.

Distribution of the Number of Bedrooms for the Houses Sampled in 2024

Number of Bedrooms	1	2	3	4	5	6
Proportion of Houses	0.12	0.22	0.28	0.22	0.14	0.02

A.

i. A house from the sample will be selected at random. What is the probability that the house had fewer than 3 bedrooms? Show your work.

ii. What is the mean number of bedrooms for the sample of newly built houses in 2024? Show your work.

B. Rodney will use a one-sample t-test for a population mean to test his belief.

i. In the context of Rodney's investigation, state the hypotheses for the test.

ii. Explain, in context, what a Type I error would be for Rodney's hypothesis test.

C. A different researcher, Keisha, suggests using a confidence interval to investigate whether the mean number of bedrooms in newly built houses in 2024 in Country B was different from 2.9.

Assume the conditions for inference have been met. Using Rodney's data, Keisha calculated a one-sample 97 percent confidence interval to estimate the population mean as $(3.01, 3.19)$. Based on the confidence interval, what conclusion can be made for Rodney's hypothesis test in part B at $\alpha = 0.03$? Justify your answer.

5. A polling agency showed the following two statements to a random sample of 1,048 adults in the United States.

Environment statement: Protection of the environment should be given priority over economic growth.

Economy statement: Economic growth should be given priority over protection of the environment.

The order in which the statements were shown was randomly selected for each person in the sample. After reading the statements, each person was asked to choose the statement that was most consistent with his or her opinion. The results are shown in the table.

	Environment Statement	Economy Statement	No Preference
Percent of sample	58%	37%	5%

(a) Assume the conditions for inference have been met. Construct and interpret a 95 percent confidence interval for the proportion of all adults in the United States who would have chosen the economy statement.

(b) One of the conditions for inference that was met is that the number who chose the economy statement and the number who did not choose the economy statement are both greater than 10. Explain why it is necessary to satisfy that condition.

(c) A suggestion was made to use a two-sample z-interval for a difference between proportions to investigate whether the difference in proportions between adults in the United States who would have chosen the environment statement and adults in the United States who would have chosen the economy statement is statistically significant. Is the two-sample z-interval for a difference between proportions an appropriate procedure to investigate the difference? Justify your answer.

4. A researcher conducted a medical study to investigate whether taking a low-dose aspirin reduces the chance of developing colon cancer. As part of the study, 1,000 adult volunteers were randomly assigned to one of two groups. Half of the volunteers were assigned to the experimental group that took a low-dose aspirin each day, and the other half were assigned to the control group that took a placebo each day. At the end of six years, 15 of the people who took the low-dose aspirin had developed colon cancer and 26 of the people who took the placebo had developed colon cancer. At the significance level $\alpha = 0.05$, do the data provide convincing statistical evidence that taking a low-dose aspirin each day would reduce the chance of developing colon cancer among all people similar to the volunteers?

Name: AP Statistics: **2015**

4. A researcher conducted a medical study to investigate whether taking a low-dose aspirin reduces the chance of developing colon cancer. As part of the study, 1,000 adult volunteers were randomly assigned to one of two groups. Half of the volunteers were assigned to the experimental group that took a low-dose aspirin each day, and the other half were assigned to the control group that took a placebo each day. At the end of six years, 15 of the people who took the low-dose aspirin had developed colon cancer and 26 of the people who took the placebo had developed colon cancer. At the significance level $\alpha = 0.05$, do the data provide convincing statistical evidence that taking a low-dose aspirin each day would reduce the chance of developing colon cancer among all people similar to the volunteers?

7 Estimating a mean – Part 1

Name: ____________ Class: ________ Date: __________

Vocabulary 词汇

Write each word three times. 把每个单词抄写三遍。

English	中文	Write it 3 times (English)
mean	均值	

Recall

In an earlier sheet you estimated a **proportion** with a confidence interval. The same idea works for an **average**:

- A sample of test scores estimates the true average score.
- The estimate comes with a margin of error.

This sheet lines up estimating a mean. Each idea has a worked example.

New ideas

1 Estimating a mean

To estimate a population **mean** 均值, take a sample, find its average, and attach a margin of error to give a confidence interval.

2 The interval

The confidence interval for a mean is sample mean ± margin of error.

Worked example. A sample of heights averages 170 cm with a margin of error of ±3 cm: the interval is 167 cm to 173 cm.

3 Interpretation

We are (say) 95% confident the true average lies in this range – the method captures it 95% of the time.

4 Precision

A larger sample gives a **smaller** margin of error, so a narrower, more precise estimate of the mean.

Watch out: the confidence interval estimates the **population mean**, not the value of a single individual. Individuals vary far more widely than the average does.

Practice

Try these in order. The methods are all above.

2.1 State what you attach to a sample mean to make a confidence interval. [1]

2.2 A sample averages 50 with a margin of error of ±4. State the confidence interval. [2]

2.3 A sample averages 200 with a margin of error of ±15. State the interval. [2]

2.4 State what a larger sample does to the margin of error. [1]

2.5 State what the interval estimates – the population mean or a single individual. [1]

Begin your response to **QUESTION 6** on this page.

SECTION II, Part B

Suggested Time—25 minutes

1 Question

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

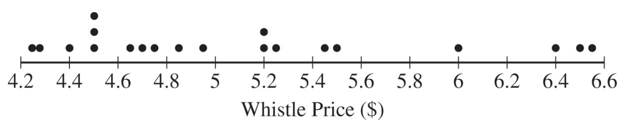
6. A company sells a certain type of whistle. The price of the whistle varies from store to store. Julio, a statistician at the company, wants to estimate the mean price, in dollars ($), of this type of whistle at all stores that sell the whistle.

(a) (i) Identify the appropriate inference procedure for Julio to use.

(ii) Describe the parameter for the inference procedure you identified in part (a-i) in context.

Julio called the managers of 20 randomly selected stores that sell the whistle and recorded the price of the whistle at each store. Following is a dotplot of Julio's data.

Price of the Whistle at 20 Stores



GO ON TO THE NEXT PAGE.

Continue your response to **QUESTION 6** on this page.

The summary statistics for Julio's data are shown in the following table.

Summary Statistics for Julio's Data

Sample Size	Mean	Standard Deviation	Minimum	Q_1	Median	Q_3	Maximum
20	5.12	0.743	4.25	4.51	4.885	5.475	6.58

(b) Julio wants to examine some characteristics of the distribution of the sample of whistle prices.

(i) Describe the shape of the distribution of the sample of whistle prices. Justify your response using appropriate values from the summary statistics table.

(ii) Using the $1.5 \times \text{IQR}$ rule, determine whether there are any outliers in the sample of whistle prices. Justify your response.

GO ON TO THE NEXT PAGE.

Continue your response to **QUESTION 6** on this page.

It can often be difficult to determine whether the distribution of sample data is skewed by looking at a graph of the data and the summary statistics, particularly when the sample size is small. Thus, statisticians sometimes measure how skewed a data set is. One such measure is Pearson's coefficient of skewness, which is calculated using the following formula.

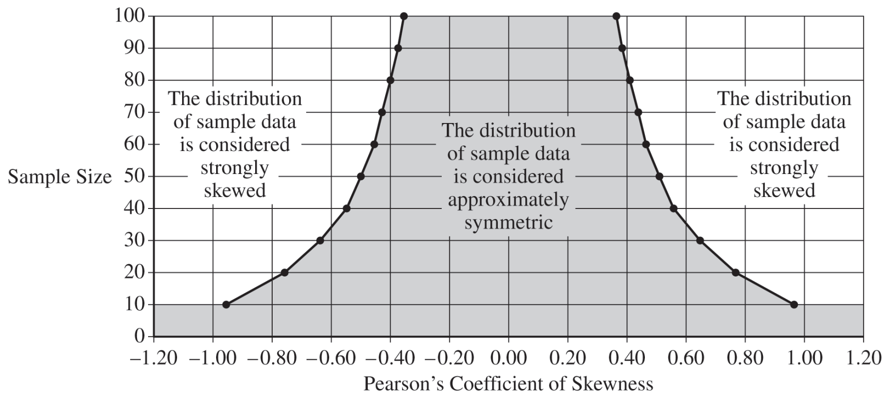
$$\text{Pearson's Coefficient of Skewness} = \frac{3(\bar{x} - m)}{s}$$

In the formula, $\bar{x}$ is the sample mean, m is the sample median, and s is the sample standard deviation.

(c) (i) Calculate Pearson's coefficient of skewness for Julio's sample of 20 whistle prices. Show your work.

The following graph shows conclusions that can be made about the shape of the distribution of sample data based on Pearson's coefficient of skewness and sample size.

Conclusion from Pearson's Coefficient of Skewness



(ii) Indicate the value of the Pearson's coefficient of skewness you calculated in part (c-i) for the appropriate sample size by marking it with an "X" on the preceding graph.

GO ON TO THE NEXT PAGE.

Continue your response to **QUESTION 6** on this page.

(d) Consider your work in part (c).

(i) What should you conclude about the shape of the distribution of the sample of whistle prices? Justify your response.

Julio's inference procedure in part (a-i) needs one of the following requirements to be satisfied to verify the normality condition.

- The sample size is greater than or equal to 30.
- If the sample size is less than 30, the distribution of the sample data is not strongly skewed and does not have outliers.

(ii) Using your response to (d-i) and the preceding requirements, is the normality condition satisfied for Julio's data? Explain your response.

GO ON TO THE NEXT PAGE.

STOP

END OF EXAM

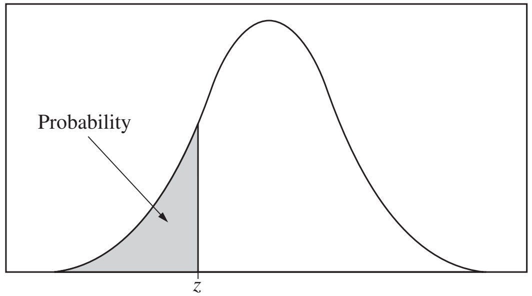


Table entry for z is the probability lying below z.

Table A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

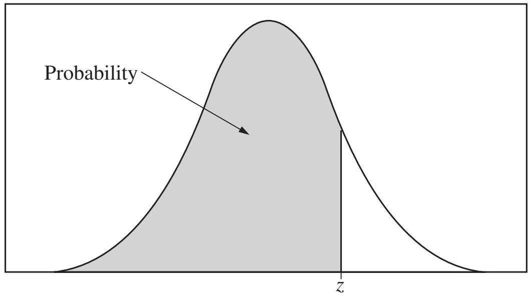


Table entry for z is the probability lying below z.

Table A ***(Continued)***

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Name: AP Statistics: **2024**

Table entry for p and C is the point t^* with probability p lying above it and probability C lying between $-t^*$ and t^*.

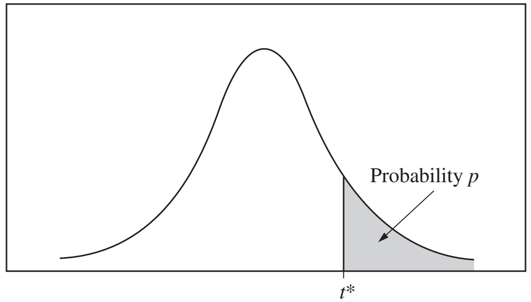


Table B ***t*** **distribution critical values**

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Table entry for p is the point (χ^2) with probability p lying above it.

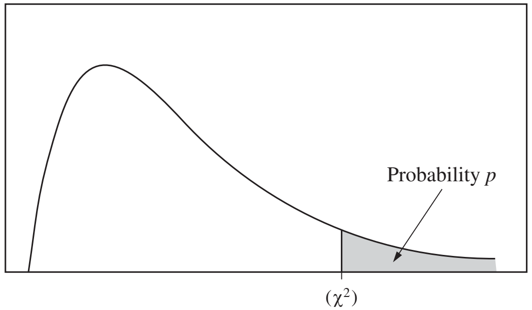


Table C χ^2 **critical values**

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.7
80	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8	128.3
100	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4	153.2

5. According to a 2017 national survey in Country B, the mean number of bedrooms in newly built houses was 2.9. Rodney, a researcher, believes the mean number of bedrooms in newly built houses in the country was different in 2024 than it was in 2017. To investigate his belief, he took a large random sample of newly built houses in Country B in 2024 and recorded the number of bedrooms in each house. The distribution of the number of bedrooms for the sampled houses is summarized in the table.

Distribution of the Number of Bedrooms for the Houses Sampled in 2024

Number of Bedrooms	1	2	3	4	5	6
Proportion of Houses	0.12	0.22	0.28	0.22	0.14	0.02

A.

i. A house from the sample will be selected at random. What is the probability that the house had fewer than 3 bedrooms? Show your work.

ii. What is the mean number of bedrooms for the sample of newly built houses in 2024? Show your work.

B. Rodney will use a one-sample t-test for a population mean to test his belief.

i. In the context of Rodney's investigation, state the hypotheses for the test.

ii. Explain, in context, what a Type I error would be for Rodney's hypothesis test.

C. A different researcher, Keisha, suggests using a confidence interval to investigate whether the mean number of bedrooms in newly built houses in 2024 in Country B was different from 2.9.

Assume the conditions for inference have been met. Using Rodney's data, Keisha calculated a one-sample 97 percent confidence interval to estimate the population mean as $(3.01, 3.19)$. Based on the confidence interval, what conclusion can be made for Rodney's hypothesis test in part B at $\alpha = 0.03$? Justify your answer.

5. According to a 2017 national survey in Country B, the mean number of bedrooms in newly built houses was 2.9. Rodney, a researcher, believes the mean number of bedrooms in newly built houses in the country was different in 2024 than it was in 2017. To investigate his belief, he took a large random sample of newly built houses in Country B in 2024 and recorded the number of bedrooms in each house. The distribution of the number of bedrooms for the sampled houses is summarized in the table.

Distribution of the Number of Bedrooms for the Houses Sampled in 2024

Number of Bedrooms	1	2	3	4	5	6
Proportion of Houses	0.12	0.22	0.28	0.22	0.14	0.02

A.

i. A house from the sample will be selected at random. What is the probability that the house had fewer than 3 bedrooms? Show your work.

ii. What is the mean number of bedrooms for the sample of newly built houses in 2024? Show your work.

B. Rodney will use a one-sample *t*-test for a population mean to test his belief.

i. In the context of Rodney's investigation, state the hypotheses for the test.

ii. Explain, in context, what a Type I error would be for Rodney's hypothesis test.

C. A different researcher, Keisha, suggests using a confidence interval to investigate whether the mean number of bedrooms in newly built houses in 2024 in Country B was different from 2.9.

Assume the conditions for inference have been met. Using Rodney's data, Keisha calculated a one-sample 97 percent confidence interval to estimate the population mean as $(3.01, 3.19)$. Based on the confidence interval, what conclusion can be made for Rodney's hypothesis test in part B at $\alpha = 0.03$? Justify your answer.

7 The t-distribution – Part 2

Name: ____________ Class: ________ Date: __________

Recall

In Part A you estimated a mean. In real life you rarely know the population's spread:

- You must estimate the spread from the sample too.
- This extra uncertainty needs a slightly different curve.

Each idea has a worked example before the Practice.

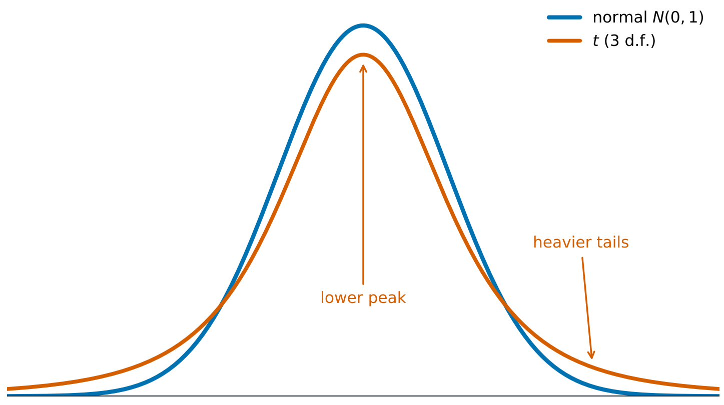
New ideas

1 When the spread is unknown

When you do not know the population's standard deviation (almost always), you estimate it from the sample. This adds uncertainty, especially for small samples.

2 The t-distribution

To handle that extra uncertainty, we use the **t-distribution** instead of the normal curve.



It looks like the bell curve but has **heavier tails**, allowing for more surprises with small samples.

■ 3 Sample size matters

The smaller the sample, the heavier the tails (more uncertainty). As the sample grows, the t-distribution gets closer and closer to the normal curve.

■ 4 Same idea, adjusted

Worked example. A confidence interval for a mean from a small sample uses a slightly wider margin (from the t-distribution) than it would with the normal curve – reflecting the extra uncertainty.

Watch out: use the **t-distribution** (not the normal) whenever you estimate the spread from the sample – which is nearly always. The two agree for large samples, but differ for small ones.

Practice

Try these in order. The methods are all above.

2.1 State why we often cannot use the population's standard deviation. [2]

2.2 State what distribution we use instead of the normal in that case. [1]

2.3 State how the t-distribution differs in shape from the normal curve. [2]

2.4 State what happens to the t-distribution as the sample size grows. [2]

2.5 State when the t-distribution should be used. [1]

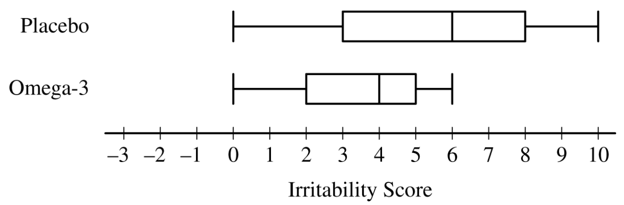
. A medical researcher completed a study comparing an omega-3 fatty acids supplement to a placebo in the treatment of irritability in patients with a certain medical condition. Nineteen patients with the medical condition volunteered to participate in the study. The study was conducted using the following weekly schedule.

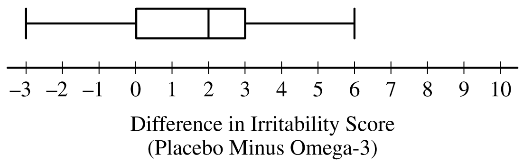
- Week 1: Each patient took a randomly assigned treatment, omega-3 supplement or placebo.
- Week 2: The patients did not take either the omega-3 supplement or the placebo. This was necessary to reduce the possibility of any carryover effect from the assigned treatment taken during week 1.
- Week 3: Each patient took the treatment, omega-3 supplement or placebo, that they did not take during week 1.

At the end of week 1 and week 3, each patient's irritability was given a score on a scale of 0 to 10, with 0 representing no irritability and 10 representing the highest level of irritability.

For each patient, the two irritability scores and the difference in their scores (placebo minus omega-3) were recorded. The results are summarized in the table and boxplots.

	n	Mean	Standard Deviation
Placebo	19	5.421	2.987
Omega-3	19	3.632	1.739
Difference (placebo minus omega-3)	19	1.789	2.485





GO ON TO THE NEXT PAGE.

Continue your response to **QUESTION 4** on this page.

The researcher claims the omega-3 supplement will decrease the mean irritability score of all patients with the medical condition similar to the volunteers who participated in the study. Is there convincing statistical evidence to support the researcher's claim at a significance level of $\alpha = 0.05$? Complete the appropriate inference procedure to support your answer.

GO ON TO THE NEXT PAGE.

4. The anterior cruciate ligament (ACL) is one of the ligaments that help stabilize the knee. Surgery is often recommended if the ACL is completely torn, and recovery time from the surgery can be lengthy. A medical center developed a new surgical procedure designed to reduce the average recovery time from the surgery. To test the effectiveness of the new procedure, a study was conducted in which 210 patients needing surgery to repair a torn ACL were randomly assigned to receive either the standard procedure or the new procedure.

   (a) Based on the design of the study, would a statistically significant result allow the medical center to conclude that the new procedure causes a reduction in recovery time compared to the standard procedure, for patients similar to those in the study? Explain your answer.

   (b) Summary statistics on the recovery times from the surgery are shown in the table.

Type of Procedure	Sample Size	Mean Recovery Time (days)	Standard Deviation Recovery Time (days)
Standard	110	217	34
New	100	186	29

   Do the data provide convincing statistical evidence that those who receive the new procedure will have less recovery time from the surgery, on average, than those who receive the standard procedure, for patients similar to those in the study?

4. The anterior cruciate ligament (ACL) is one of the ligaments that help stabilize the knee. Surgery is often recommended if the ACL is completely torn, and recovery time from the surgery can be lengthy. A medical center developed a new surgical procedure designed to reduce the average recovery time from the surgery. To test the effectiveness of the new procedure, a study was conducted in which 210 patients needing surgery to repair a torn ACL were randomly assigned to receive either the standard procedure or the new procedure.

(a) Based on the design of the study, would a statistically significant result allow the medical center to conclude that the new procedure causes a reduction in recovery time compared to the standard procedure, for patients similar to those in the study? Explain your answer.

(b) Summary statistics on the recovery times from the surgery are shown in the table.

Type of Procedure	Sample Size	Mean Recovery Time (days)	Standard Deviation Recovery Time (days)
Standard	110	217	34
New	100	186	29

Do the data provide convincing statistical evidence that those who receive the new procedure will have less recovery time from the surgery, on average, than those who receive the standard procedure, for patients similar to those in the study?

8 Categorical data and expected counts – Part 1

Name: ____________ Class: ________ Date: __________

Vocabulary 词汇

Write each word three times. 把每个单词抄写三遍。

English	中文	Write it 3 times (English)
Categorical data	分类数据	
expected	期望	

Recall

Not all data are numbers – some are categories:

- Eye colour, favourite subject, or blood type are categories.
- You count how many fall into each category.

This sheet lines up how to compare category counts. Each idea has a worked example.

New ideas

1 Categorical data

Categorical data 分类数据 sort things into groups (categories) rather than measuring a number. You summarize them with **counts** in each category.

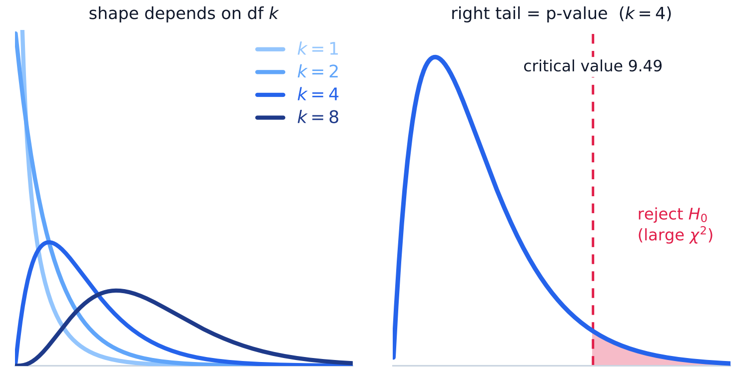
2 Observed and expected

An **observed** count is what you actually got; an **expected** 期望 count is what you would expect if a claim were true.

Worked example. Rolling a fair die 60 times, you **expect** 10 of each number. If you **observe** 15 sixes, that is more than expected.

3 Measuring the gap

To judge whether the observed counts differ from expected by more than chance, we compare each observed count with its expected count.



$\chi^2 = \sum \frac{(O-E)^2}{E}$ · large values reject H_0 (right-tail only)

4 The question

The key question: are the differences between observed and expected small enough to be chance, or big enough to be real?

Watch out: the **expected** counts are what a claim (like "the die is fair") predicts – they are not the observed data. The whole test compares what you **got** against what the claim **predicts**.

Practice

Try these in order. The methods are all above.

2.1 State what categorical data are, with an example. [2]

2.2 State the difference between an observed and an expected count. [2]

2.3 A fair coin is flipped 100 times. State the expected number of heads. [2]

2.4 A fair die is rolled 120 times. State the expected count for each number. [2]

2.5 State the key question when comparing observed and expected counts. [2]

2. Product advertisers studied the effects of television ads on children's choices for two new snacks. The advertisers used two 30-second television ads in an experiment. One ad was for a new sugary snack called Choco-Zuties, and the other ad was for a new healthy snack called Apple-Zuties.

For the experiment, 75 children were randomly assigned to one of three groups, A, B, or C. Each child individually watched a 30-minute television program that was interrupted for 5 minutes of advertising. The advertising was the same for each group with the following exceptions.

- The advertising for group A included the Choco-Zuties ad but not the Apple-Zuties ad.
- The advertising for group B included the Apple-Zuties ad but not the Choco-Zuties ad.
- The advertising for group C included neither the Choco-Zuties ad nor the Apple-Zuties ad.

After the program, the children were offered a choice between the two snacks. The table below summarizes their choices.

Group	Type of Ad	Number Who Chose Choco-Zuties	Number Who Chose Apple-Zuties
A	Choco-Zuties only	21	4
B	Apple-Zuties only	13	12
C	Neither	22	3

(a) Do the data provide convincing statistical evidence that there is an association between type of ad and children's choice of snack among all children similar to those who participated in the experiment?

(b) Write a few sentences describing the effect of each ad on children's choice of snack.

8 The chi-square test – Part 2

Name: ___________ Class: ________ Date: __________

Vocabulary 词汇

Write each word three times. 把每个单词抄写三遍。

English	中文	Write it 3 times (English)
chi-square	卡方	

Recall

In Part A you compared observed counts with expected counts. Now measure the total gap with one number:

- We need a single value for "how far off" the observed counts are.
- That value is the chi-square statistic.

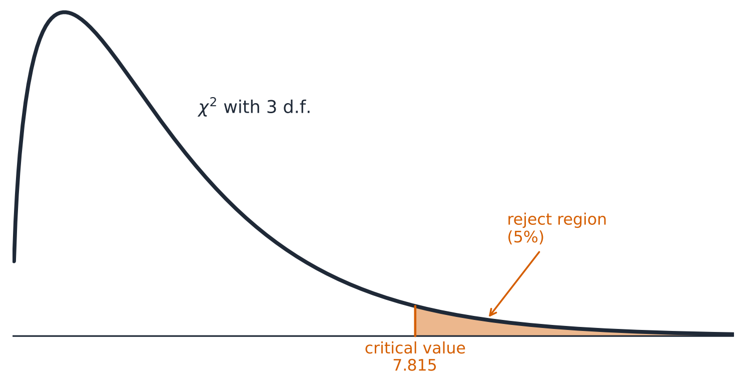
Each idea has a worked example before the Practice.

New ideas

1 The chi-square statistic

The **chi-square** 卡方 statistic adds up the (observed minus expected) gaps, squared and scaled:

$$\chi^2 = \sum \frac{(O-E)^2}{E}.$$



A **bigger** χ^2 means the observed counts are further from expected.

2 A worked example

Worked example. A cross predicts 60 dominant and 20 recessive, but you observe 55 and 25:

$$\chi^2 = \frac{(55-60)^2}{60} + \frac{(25-20)^2}{20} = \frac{25}{60} + \frac{25}{20} = 0.42 + 1.25 = 1.67.$$

3 Making a decision

Compare χ^2 with a **critical value** from a table. If χ^2 is **smaller**, the differences are within chance (do not reject); if **larger**, they are a real effect (reject).

4 In the example

With 1 degree of freedom the critical value at the 0.05 level is 3.84. Since $1.67 < 3.84$, the deviation is within chance – the data fit the predicted ratio.

Watch out: always divide by the **expected** count E (not the observed) in each term. And a small χ^2 means the data **fit** the claim – it does not mean "no data".

Practice

Try these in order. The methods are all above.

2.1 Write the formula for the chi-square statistic. [2]

2.2 State what a large chi-square value tells you. [2]

2.3 For one category, observed $= 18$ and expected $= 12$. Find its term $\dfrac{(O-E)^2}{E}$. [3]

2.4 A test gives $\chi^2 = 1.2$ and the critical value is 3.84. State the decision. [2]

2.5 State what you divide by in each term of the chi-square sum. [1]

2. Product advertisers studied the effects of television ads on children's choices for two new snacks. The advertisers used two 30-second television ads in an experiment. One ad was for a new sugary snack called Choco-Zuties, and the other ad was for a new healthy snack called Apple-Zuties.

   For the experiment, 75 children were randomly assigned to one of three groups, A, B, or C. Each child individually watched a 30-minute television program that was interrupted for 5 minutes of advertising. The advertising was the same for each group with the following exceptions.

   - The advertising for group A included the Choco-Zuties ad but not the Apple-Zuties ad.
   - The advertising for group B included the Apple-Zuties ad but not the Choco-Zuties ad.
   - The advertising for group C included neither the Choco-Zuties ad nor the Apple-Zuties ad.

   After the program, the children were offered a choice between the two snacks. The table below summarizes their choices.

Group	Type of Ad	Number Who Chose Choco-Zuties	Number Who Chose Apple-Zuties
A	Choco-Zuties only	21	4
B	Apple-Zuties only	13	12
C	Neither	22	3

   (a) Do the data provide convincing statistical evidence that there is an association between type of ad and children's choice of snack among all children similar to those who participated in the experiment?

   (b) Write a few sentences describing the effect of each ad on children's choice of snack.

9 Inference for a regression slope – Part 1

Name: ____________ Class: ________ Date: __________

Vocabulary 词汇

Write each word three times. 把每个单词抄写三遍。

English	中文	Write it 3 times (English)
slope	斜率	

Recall

Earlier you fitted a line of best fit to a scatterplot:

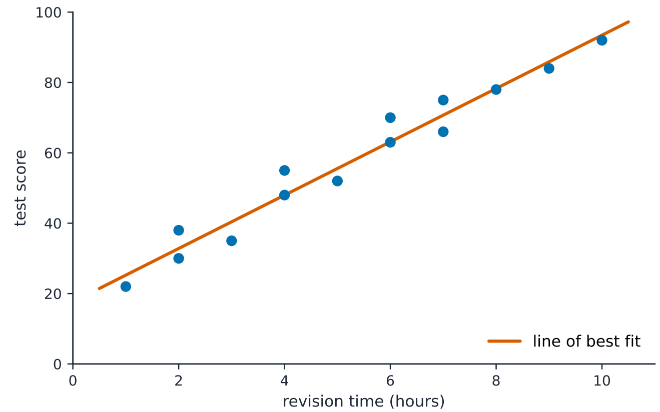
- The line's slope describes the relationship.
- But is that slope real, or could it just be chance in the sample?

This sheet lines up inference about a regression line. Each idea has a worked example.

New ideas

■ 1 The sample slope estimates the true slope

The **slope** 斜率 of the line of best fit, found from a sample, is only an **estimate** of the true relationship in the whole population.



■ 2 Is there a real relationship?

If the true slope is 0, there is **no** linear relationship. So we test whether the slope is really different from 0.

■ 3 A confidence interval for the slope

We can also give a confidence interval for the slope – a range likely to contain the true slope.

Worked example. If a sample slope is 5 with a margin of error of ± 2, the interval is 3 to 7. Since this does not include 0, there is evidence of a real positive relationship.

■ 4 The idea reused

This is the same inference idea as before: a sample estimate, plus uncertainty, used to judge a claim (here, "the slope is 0").

Watch out: if a slope's confidence interval **includes** 0, you cannot conclude there is a real relationship – the true slope might be 0 (no link at all).

Practice

Try these in order. The methods are all above.

2.1 State what the sample slope is an estimate of. [2]

2.2 State what a true slope of 0 would mean. [2]

2.3 A sample slope is 8 with a margin of error of ± 3. State the confidence interval. [2]

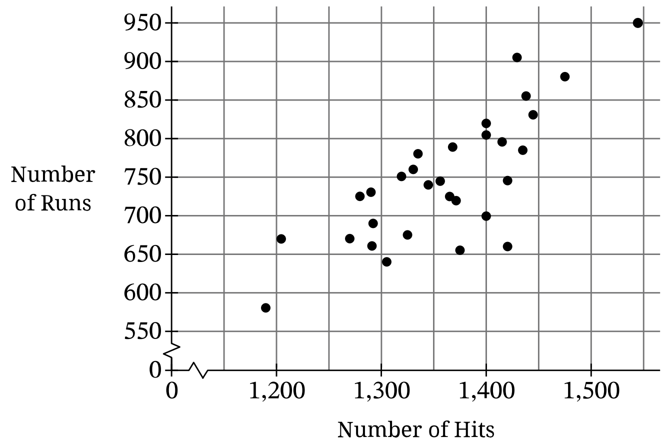
2.4 For that interval, state whether there is evidence of a real relationship, and why. [2]

2.5 State what it means if a slope's interval includes 0. [1]

The following information applies to parts A and B.

6. To win a game, a baseball team needs to score runs. To score runs, players need to get on base. The primary way players get on base is by hitting the ball. Figure 1 shows the relationship between the number of hits and the number of runs for 30 randomly selected professional baseball teams.

Figure 1. Scatterplot of Number of Runs by Number of Hits



Part A

Consider the scatterplot in Figure 1.

i. Describe the relationship between the number of hits and the number of runs for these baseball teams, in context.

ii. Based on the data from the sample of 30 teams, an equation of the least-squares regression line model for predicting the number of runs from the number of hits is as follows.

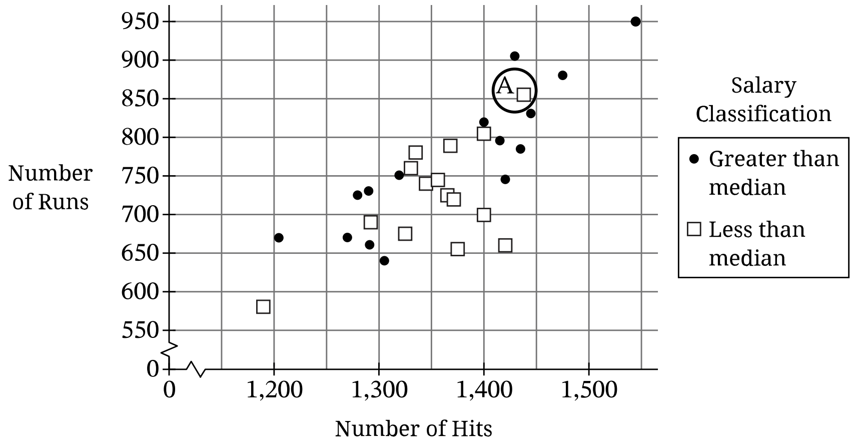
$$\text{Predicted number of runs} = -372.2 + 0.823(\text{number of hits})$$

Using the given regression equation, calculate the point estimate for the predicted number of runs a team would score if they were to achieve their goal of 1,250 hits. Show your work.

Part B

Each team has a total salary equal to the sum of the salaries of all players on the team. The median total salary for the sample of 30 teams is $160 million. In Figure 2, points with dots represent teams with a total salary greater than the median, and points with squares represent teams with a total salary less than the median.

Figure 2. Scatterplot of Number of Runs by Number of Hits and Salary Classification



Consider the scatterplot in Figure 2.

i. Compare the team represented by the point that is circled and labeled A with the other teams that have the same total salary classification.

ii. For each salary classification, consider the linear relationship between the number of hits and the number of runs. For teams with a total salary greater than the median, is the strength of the linear relationship stronger than, weaker than, or similar to the strength of the linear relationship for teams with a total salary less than the median? Explain.

Part C

Different types of intervals can be used when working with linear regression models. One type of interval is a confidence interval for the slope of the least-squares regression line. Two other types of intervals in the context of this problem are as follows:

- A confidence interval that estimates the mean number of runs for all teams with a specific number of hits.
- A prediction interval that predicts the number of runs for a single team with a specific number of hits.

All three intervals use the same basic formula:

$$\text{Point Estimate} \pm t^{*}(\text{standard error}),$$

where critical value t^* comes from a t-distribution that has $n-2$ degrees of freedom. The same critical value is used when finding each of these intervals with the same level of confidence.

Consider the point estimate for the predicted number of runs found in part A (ii) for the team whose goal is to achieve 1,250 hits next year. Recall that this value was calculated using the linear regression model based on the data from the sample of 30 teams.

i. What is the critical value for a 95% confidence level that would be used for the confidence interval for the mean number of runs as well as for the prediction interval for the number of runs? Indicate the answer to two decimal places.

ii. The standard error for the confidence interval for the mean number of runs is 17.48. Assuming the conditions for inference are met, calculate the 95% confidence interval for the mean number of runs for all teams with 1,250 hits. Show your work.

iii. The standard error for the prediction interval for the number of runs is 56.78. Assuming the conditions for inference are met, calculate the 95% prediction interval for the number of runs for a single team with 1,250 hits. Show your work.

Part D

i. Would a distribution of sample means be expected to have more variability or less variability than a distribution of individual observations? Explain.

ii. The standard error used in the confidence interval that estimates the mean number of runs for all teams with x hits can be found using the following formula.

$$\sqrt{s^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

The standard error used in the prediction interval that predicts the number of runs for a single team with x hits can be found using the following formula.

$$\sqrt{s^2+s^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

In both standard error formulas, s is the standard deviation of the residuals, n is the sample size, and $\bar{x}$ is the sample mean number of hits. Based on the answer from part D (i) and the standard error formulas for the confidence interval and the prediction interval, explain why the prediction interval calculated in part C (iii) is wider than the confidence interval calculated in part C (ii).

STOP

END OF EXAM

AP STATISTICS • HOLIDAY HOMEWORK 假期作业

9 Residuals and checking the model – Part 2

Name: ____________ Class: ________ Date: __________

Vocabulary 词汇

Write each word three times. 把每个单词抄写三遍。

English	中文	Write it 3 times (English)
residual	残差	____________ ____________ ____________

Recall

In Part A you tested a regression slope. Now check whether a line is even the **right** model:

- A straight line is not always the best fit.
- The leftover errors (residuals) reveal whether it fits.

Each idea has a worked example before the Practice.

New ideas

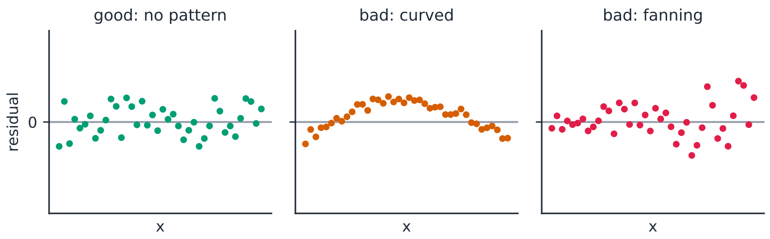
■ 1 Residuals

A **residual** 残差 is the gap between an actual value and the line's prediction: residual = actual − predicted.

Worked example. If the line predicts 70 but the actual value is 73, the residual is 73 − 70 = +3.

■ 2 The residual plot

A **residual plot** graphs the residuals against x. It magnifies how the line misses the points.



residual $= y - \hat{y}$ · random scatter about 0 supports a linear model (LINER)

■ **3 No pattern is good**

If the residual plot shows **no pattern** (a random scatter around zero), a straight line is a good model.

■ **4 A pattern is bad**

If the residual plot shows a **curve** or other pattern, a straight line is the **wrong** model – the data bend in a way the line cannot follow.

Watch out: a good model has a residual plot with **no pattern**. A clear pattern (like a U-shape) is a warning that a straight line is the wrong choice, even if the correlation looked high.

Practice

Try these in order. The methods are all above.

2.1 State what a residual is. [2]

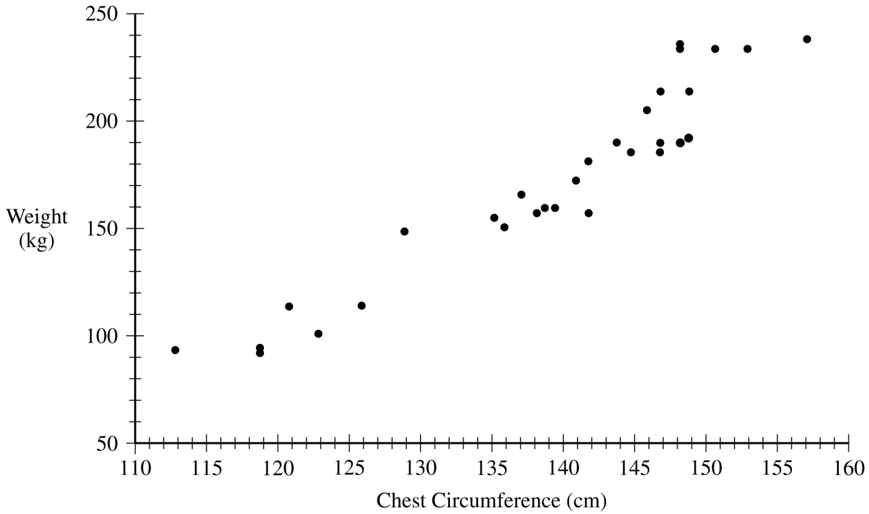
2.2 A line predicts 50 but the actual value is 46. Find the residual. [2]

2.3 State what a residual plot graphs. [2]

2.4 State what a residual plot with no pattern tells you. [2]

2.5 State what a clear pattern (like a curve) in the residual plot tells you. [2]

. Wildlife biologists are interested in the health of tule elk, a species of deer found in California. An important measurement of tule elk health is their weight. The weight of a tule elk is difficult to measure in the wild. However, chest circumference, which is believed to be related to the weight of a tule elk, can easily be measured from a safe distance using a harmless laser. A study was done to investigate whether chest circumference, in centimeters (cm), could be used to accurately estimate the weight, in kilograms (kg), of male tule elk. For the study, wildlife biologists captured 30 male tule elk, measured their chest circumference and weight, and then released the elk. The data for the 30 male tule elk are shown in the scatterplot.



(a) Describe the relationship between chest circumference and weight of male tule elk in context.

GO ON TO THE NEXT PAGE.

Continue your response to **QUESTION 5** on this page.

Following is the equation of the least-squares regression line relating chest circumference and weight for male tule elk.

$$\text{predicted weight} = -350.3 + 3.7455(\text{chest circumference})$$

(b) The weight of one male tule elk with a chest circumference of 145.9 cm is 204.3 kg.

(i) Using the equation of the least-squares regression line, calculate the predicted weight for this male tule elk. Show your work.

(ii) Calculate the residual for this male tule elk. Show your work.

GO ON TO THE NEXT PAGE.

Continue your response to **QUESTION 5** on this page.

The equation of the least-squares regression line relating chest circumference and weight for male tule elk is repeated here.

$$\text{predicted weight} = -350.3 + 3.7455(\text{chest circumference})$$

(c) Interpret the slope of the least-squares regression line in context.

(d) The sambar, another species of deer, is similar in size to the tule elk. The slope of the population regression line relating chest circumference and weight for all male sambars is 4.5 kilograms per centimeter. A wildlife biologist wants to determine whether the slope of the population regression line for male tule elk is different than that for male sambars. Let β represent the slope of the population regression line for male tule elk. The wildlife biologist conducted a test of the following hypotheses using the sample of 30 tule elk.

$$H_0: \beta = 4.5$$
$$H_a: \beta \neq 4.5$$

The test statistic was calculated to be 3.408. Assume all conditions for inference were met.

(i) Determine the p-value of the test.

GO ON TO THE NEXT PAGE.

Continue your response to **QUESTION 5** on this page.

(ii) At a significance level of $\alpha = 0.05$, what conclusion should the wildlife biologist make regarding the slope of the population regression line for male tule elk? Justify your response.

GO ON TO THE NEXT PAGE.