12TH EDITION

STATISTICS AT SQUARE ONE

Edited by
MICHAEL J. CAMPBELL

with website

WILEY Blackwell

Table of Contents

- Cover
- Title Page
- Copyright Page
- Dedication Page
- Preface
- About the companion website
- CHAPTER 1: Understanding basic numbers
  - When is a number large?
  - Ratios
  - Using ratios to adjust for other variables
  - Proportions, percentages and odds
  - Percentage difference and percentage change: importance of baseline
  - Rounding proportions and percentages
  - Probabilities and risks
  - Prevalence and incidence rate
  - Trusting numbers
  - Conclusions
  - Further reading
  - Exercises
  - References
- CHAPTER 2: Data display and summary
  - Types of data
  - Stem-and-leaf plots and dot plots
  - Box-whisker plots
  - Frequency tables and histograms

- Bar charts
- Further reading
- Common questions
- Displaying data in papers
- References

CHAPTER 3: Summary statistics for quantitative data

- Mean
- Variance and standard deviation
- Normal distribution
- Skewness
- Between-subjects and within-subjects standard deviation
- Common questions
- Formula appreciation
- Reading and Displaying Summary Statistics
- References

CHAPTER 4: Summary statistics for binary data

- Summarising one binary variable
- Summarising the relationship between two binary variables
- Odds ratios and cross-sectional studies
- Odds ratios and case–control studies
- Common questions
- Reading and displaying summary statistics
- References

CHAPTER 5: Diagnostic and screening tests

- Diagnostic and screening tests
- Examples
- Sensitivity and Specificity

Positive predictive value in relation to prevalence
Likelihood ratio
Receiver operating characteristics curves
Further discussion on diagnostic and screening tests
Limitations of the conventional diagnostic testing paradigm
Reading and reporting diagnostic/screening tests
Exercises
References
CHAPTER 6: Populations and samples
Populations
Samples
Unbiasedness and precision
Problems of bias in non-randomised samples (especially Big Data)
Randomisation
Variation between samples
Standard error of the mean
Example of standard error
Standard error of a proportion or a percentage
Problems with non-random samples
Common questions
Important points
Reading and reporting populations and samples
Exercises
References
CHAPTER 7: Statements of probability and confidence intervals

Reference ranges
Confidence intervals
Large sample standard error of difference between means
Large sample confidence interval for the difference in two means
Standard error of difference between percentages or proportions
Confidence interval for a difference in proportions or percentages
Confidence interval for an odds ratio
Confidence interval for a relative risk
Confidence Intervals for other estimates
Common Questions
Reading and reporting confidence intervals
Formula appreciation
Exercises
References
CHAPTER 8: P values, power, type I and type II errors
Null hypothesis and type I error
Testing for differences of two means
Testing for a difference in two proportions
P value
P values, confidence intervals and clinically important results
Alternative hypothesis and type II error
Other types of statistical inference
Issues with P values
One-sided and two-sided tests

Tests for superiority, tests for non-inferiority and tests for equivalence
Links with diagnostic tests
Common questions
Reading and reporting significance tests
References
CHAPTER 9: Tests for differences between two groups of a quantitative outcome with small samples
Student’s *t* test
Confidence interval for the mean froma small sample
Difference of sample mean from population mean (one-sample *t* test)
Difference between means of two samples
Unequal standard deviations
Difference between means of paired samples (paired *t* test)
Non-parametric or distribution-free tests
Tests for differences in unpaired samples of non-Normally distributed data (Mann–Whitney U test)
Tests for differences in paired samples of non-Normally distributed data (Wilcoxon test)
Computer-intensive methods
Discussion
Reading and reporting *t* tests and non-parametric tests
Common questions
References
CHAPTER 10: Tests for association in binary and categorical data

General chi-squared test
2 × 2 tables
χ^2 test for trend
Comparison of an observed and a theoretical distribution
Tests for paired binary data
Examples of a paired comparison
Extensions of the χ^2 test
Common questions
Formula appreciation
Reading and reporting chi-squared tests
References
CHAPTER 11: Correlation and regression
The correlation coefficient
Looking at data: scatter diagrams
Calculation of the correlation coefficient
Significance test for a correlation coefficient
Spearman rank correlation
The regression equation
Simple checks of the model
Using regression in *t* tests
More advanced methods
Common questions
Formula appreciation
Reading and reporting correlation and regression
References
CHAPTER 12: Survival analysis
Why survival analysis is different

Kaplan–Meier survival curve
Example of calculation of survival curve
The log rank test
Further methods
Common questions
Reading and reporting survival analysis
References
CHAPTER 13: Modelling data
Basics
Models
Model fitting and analysis: exploratory and confirmatory analyses
Bayesian methods
Models generally
Multiple linear regression
Example linear regression
Logistic regression
Survival analysis
Other things to consider in modelling
References
CHAPTER 14: Study design and choosing a statistical test
Design
Sample size
Choice of test
Reading and reporting on the design of a study
Further reading
References
CHAPTER 15: Use of computer software

Chapter 2: Data display and summary
Chapter 3: Summary statistics for quantitative data
Chapter 4: Summary statistics for binary data
Chapter 5: Diagnostic and screening tests
Chapter 6: Populations and samples
Chapter 7: Statements of probability and confidence intervals
Chapter 8: P values, power, type I and type II errors
Chapter 9: Tests for differences between two groups of a quantitative outcome with small samples
Chapter 10: Tests for association in binary and categorical data
Chapter 11: Correlation and regression
Chapter 12: Survival analysis
Chapter 13: Modelling data
Appendix
Software packages
References
Answers to exercises
Glossary of statistical terms
Appendix
Index
End User License Agreement

List of Tables

Chapter 2
Table 2.1 Examples of types of data.

Table 2.2 Urinary concentration of lead in 15 children from a housing estate ...

Table 2.3 Urinary concentration of lead in 16 rural children (μmol/24 h).

Table 2.4 Concentrations of urinary copper (in μmol/24 h) in 40 children.

Table 2.5 Lead concentration in 140 urban children.

Chapter 3

Table 3.1 Calculation of standard deviation for ungrouped data from Table 2.2...

Table 3.2 Number of examinations by a doctor in the past year.

Chapter 4

Table 4.1 2 × 2 contingency table for comparison of two groups

Table 4.2 Results from the isoniazid trial after six months' follow-up

Table 4.3 Odds ratios and relative risks for different values of absolute ris...

Table 4.4 Results from the PHVD trial in 151 premature babies

Table 4.5 2 × 2 table for association studies

Table 4.6 Association between hay fever and eczema in children aged 11

Table 4.7 2 × 2 table for a case–control study

Table 4.8 Outcome of a case–control study of the risk of a coronary artery an...

Table 4.9 2 × 2 table for a case–control study with double the number of cont...

Chapter 5

Table 5.1 Standard table for diagnostic and screening tests.

Table 5.2 Results of FebriDx (test 1) in 248 patients with suspected COVID-19

Table 5.3 Results in 965 patients attending primary care.

Table 5.4 Results of a diagnostic test with 100% sensitivity.

Table 5.5 Results of a diagnostic test with 100% specificity

Table 5.6 Standard situation but with a doubling of the prevalence

Chapter 6

Table 6.1 Mean diastolic blood pressures of employed and unemployed men aged ...

Chapter 7

Table 7.1 Appendicitis data

Chapter 8

Table 8.1 Relationship between type I and type II errors.

Chapter 9

Table 9.1 Transit times of marker pellets through the alimentary canal of pat...

Table 9.2 Transit times of marker pellets through the alimentary canal of 12 ...

Table 9.3 Plasma globulin fraction after randomisation to treatment A or B.

Table 9.4 Combined results of Table 9.3.

Table 9.5 Wilcoxon test on foetal movement before and after chorionic villus ...

Table 9.6 P values for different tests for Table 9.1.

Table 9.7 Confidence intervals (CIs) for differences in means associated with...

Chapter 10

Table 10.1 Distribution by city suburb of patients admitted to self-poisoning...

Table 10.2 Calculation of the χ^2 test on figures in Table 10.1.

Table 10.3 Results from the isoniazid trial after 6 months' follow-up.

Table 10.4 Change in eating poultry in randomised controlled trial.

Table 10.5 Calculation of X^2 for comparison between actual distribution and t...

Table 10.6 Results from a matched or paired study.

Table 10.7 Layout for paired data.

Table 10.8 Data on 539 subjects in the active treatment group in the REGASSA ...

Chapter 11

Table 11.1 Height and pulmonary anatomical dead space in 15 children.

Table 11.2 Derivation of Spearman rank correlation from data in Table 11.1.

Chapter 12

Table 12.1 Survival in 49 patients with Dukes' C colorectal cancer randomly a...

Table 12.2 Calculation of survival case for 25 patients randomly assigned to ...

Table 12.3 Calculation of log rank statistics for 49 patients randomly assign...

Chapter 13

Table 13.1 Possible values for the expected value E.

Table 13.2 Predictors of sitting time in a convenience survey (Dunton).

Table 13.3 Factors related to myopia in China (selected logistic regression c...

Table 13.4 Partial results of a Cox regression Outcome: 1 year mortality, tot...

Chapter 14

Table 14.1 Choice of statistical test from paired or matched observations

Appendix

Table A Probabilities related to multiples of standard deviations for a Normal d...

List of Illustrations

Chapter 2

Figure 2.1 Stem and leaf 'as they come' from Table 2.2.

Figure 2.2 Ordered stem-and-leaf plot from Table 2.2.

Figure 2.3 Dot plot of urinary lead concentration for urban and rural childr...

Figure 2.4 Box-whisker pot of data from Tables 2.2 and 2.3.

Figure 2.5 Box-whisker plot of data from Table 2.4.

Figure 2.6 Histogram of the data on lead concentration in 140 children from ...

Figure 2.7 Bar chart of housing data for 140 children and comparable census ...

Chapter 3

Figure 3.1 Normal curve and histogram of diastolic blood pressure from 500 m...

Figure 3.2 Distribution of disposable income in the UK in 2019 showing the d...

Chapter 5

Figure 5.1 Line diagram showing the outcome from testing 10,000 people with ...

Figure 5.2 Line diagram showing the outcome from testing 10,000 people with ...

Figure 5.3 A receiver operator characteristics curve for different cut-offs ...

Chapter 6

Figure 6.1 An illustration of bias and precision.

Figure 6.2 Demonstration of the central limit theorem using *cltdemo*.

Chapter 7

Figure 7.1 100 simulated confidence intervals with mean 88 and standard devi...

Chapter 8

Figure 8.1 95% confidence intervals for five studies.

Figure 8.2 The relationship between type 1 error rate α, type II error rate ...

Chapter 9

Figure 9.1 Dot-plot of transit time data

Figure 9.2 Data on plasma globulin fraction from Table 9.3.

Figure 9.3 Difference in foetal movement from Table 9.5.

Chapter 11

Figure 11.1 Types of relationship.

Figure 11.2 Scatter plot of dead space versus height from Table 11.1.

Figure 11.3 The regression line $Y = \beta_0 + \beta_1 x$.

Figure 11.4 Regression line drawn on scatter diagram relating height and pul...

Figure 11.5 Plot of the residuals from the regression line applied to dead s...

Figure 11.6 A Q–Q plot of the residuals from the dead space data.

Chapter 12

Figure 12.1 Illustration of censored observations. × = event, + = censored e...

Figure 12.2 Survival curves against time (months) of 25 patients with Dukes'...

Chapter 13

Figure 13.1 A quadratic line $y = 1 – x − x^2$.

Figure 13.2 Linear regression when X_2 is binary.

Figure 13.3 Linear regression including an interaction between X_1 and X_2.

Chapter 14

Figure 14.1 Illustration of power using confidence intervals (CIs).

Figure 14.2 Choosing a test to compare two independent groups.

Statistics at Square One

Twelfth Edition

Edited by

Michael J. Campbell
Emeritus Professor of Medical Statistics
Medical Statistics Group, School of Health and Related Research
University of Sheffield
Sheffield, UK

WILEY Blackwell

This edition first published 2021

Edition History
M.J. Campbell and T.D.V. Swinscow (11e, 2009). BMJ Books was an imprint of BMJ Publishing Group Limited, used under licence by Blackwell Publishing which was acquired by John Wiley & Sons in February 2007. Previous editions: 1976, 1977, 1978, 1978, 1979, 1980, 1980, 1983, 1996, 2002

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
9600 Garsington Road, Oxford, OX4 2DQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: Campbell, Michael J., 1950– author.
Title: Statistics at square one / Michael J. Campbell.
Description: Twelfth edition. | Hoboken, NJ : Wiley-Blackwell, 2021. | Includes bibliographical references and index.
Identifiers: LCCN 2021019285 (print) | LCCN 2021019286 (ebook) | ISBN 9781119401308 (paperback) | ISBN 9781119401421 (adobe pdf) | ISBN 9781119402343 (epub)
Subjects: MESH: Statistics as Topic Classification: LCC RA407 (print) | LCC RA407 (ebook) | NLM WA 950 | DDC 610.72/7–dc23
LC record available at https://lccn.loc.gov/2021019285
LC ebook record available at https://lccn.loc.gov/2021019286

Cover Design: Wiley
Cover Image: © Somatuscani/iStock/Getty Images

To Matthew, Annabel, Chloe, Robyn, Charlie, Flora and Edith.

Preface

This book is aimed at anyone who needs a basic introduction to statistics in the health sciences. It is based on many years' experience teaching first-year medical and health science students. Many of the examples are taken from primary care in the UK, which is where I worked for many years. Throughout I have tried to emphasise that medical statistics is not just a bag of tricks, and there are many synergies between its methods.

It is now over 40 years since Swinscow's original edition of this book, and each edition reflected changes in the understanding of medical statistics. Perhaps the greatest change has occurred since the previous edition, which appeared 12 years ago. Despite the efforts of medical statisticians, there was a widespread misuse of P values, the cornerstone of conventional statistical inference. This led some journals to ban their use altogether. It is my view that used properly the P value is a useful concept, but this book, as in previous editions, concentrates on estimation rather than just hypothesis testing. The book tries to steer the reader away from an excessive devotion to P values, to instil a proper appreciation of their usefulness and to emphasise estimation over significance testing.

This book was revised during the COVID-19 pandemic, which drew attention to the usefulness of statistics to understand public health and so there are a number of COVID-related examples. One area where there has been much attention is the use of diagnostic tests and the relevant chapter has been considerably updated in light of the pandemic.

There have been other important changes in the statistical arena since the 11th edition. Free statistical software has become more generally available and is easier to use, particularly R with RStudio and R commander, so I have rewritten all the examples and figures in that package. All the code is given, making replication easy. The package OpenEpi remains useful and very easy to use, so I have retained some examples applying it. Computer-intensive methods such as the bootstrap are readily understood and now easily implemented, so they are included. The links between methods are described, and this is made easier with computer-intensive methods, which do not require specific assumptions for different methods. The formulas and worked examples are retained because without them the computer software is just a ‘black box’. The exercises on ‘playing with the data’ are also retained, since the advantage of using computers is that it is little additional effort to change the data and see the effect on the results. This kind of exercise emphasises which assumptions are important and which are less so.

This 12th edition comes with two new chapters. The first is on understanding basic numbers. This may seem somewhat elementary, but it has been my experience that many newspapers and politicians misuse basic data, to such an extent that the misuse is often accepted without comment, so I hope this chapter will provide a handy guide to scepticism on official pronouncements. I have also added a new chapter on modelling. Even new students will have to read the current literature and most papers in the health science literature now use models, so an appreciation of their use and misuse is required. For greater depth I refer the reader to a companion book to this one, Walters *et al.*’s *Medical Statistics*.[1] In addition, we have published a checklist that we hope will prove helpful for students struggling to interpret the statistics of a published paper.[2]

Feedback from previous editions has indicated that the Commonly Asked Questions are a useful critique of the methods. As before, each chapter contains exercises, some of which are based on the Royal College of General Practitioners' (RCGP) Advanced Knowledge Test. There are answers to these exercises at the back of the book.

I am grateful to my colleagues Stephen Walters, Nigel Mathers and Dan Green who kindly commented on various parts of this book, to Pete Dodd who helped put the R programs on Github (https://github.com/mikejcampbell50/StatsSq1) and to Daniel Barker of the University of Newcastle, New South Wales for comments on Chapter 1. I am grateful to them and to my former colleagues Steven Julious, Richard Jacques and Dawn Teare, for support and from whom I learnt a great deal.

MJ Campbell
Sheffield, UK

1. Walters SJ, Campbell MJ, Machin D . *Medical statistics: A textbook for the health sciences* , 5th edn. Chichester: Wiley, 2020.

2. Mansournia MA *et al*. A CHecklist for statistical Assessment of Medical Papers (the CHAMP statement): explanation and elaboration. *Br J Sports Med*. 2021. doi:10.1136/ bjsports-2020-103652.

About the companion website

The companion website contains all the R programs in the book. They can be copied electronically and can be used for teaching and to perform statistical tests.

http://www.wiley.com/go/Campbell12e

CHAPTER 1
Understanding basic numbers

Numbers are not necessarily easy to understand and, notwithstanding stories of grandmothers and teaching them to suck eggs, this chapter will try to cover some of the basics for understanding numbers. The chapter warns about ‘orphan’ numbers and how percentage changes are difficult.

When is a number large?

Consider the following examples:

1. On 6 May 2020 there were 30,000 deaths due to COVID-19 in the UK, 75,000 in the USA and 265,000 in the world.[1]
2. There were about 634,000 deaths in the UK population, 2,909,000 in the USA and 58 million deaths in the world in 2018.[2]
3. The UK Government stated in 2018: ‘We have invested an extra £1 billion in the NHS [National Health Service] this year.’[3]
4. The UK sent £350 million to the European Union every week.[4]
5. The Global Burden of Disease Report (GBDR) on sepsis estimated that there were 48.9 million cases in 2017, and 11 million deaths, across 195 countries and territories.[5]

Are these large amounts? They certainly sound like large amounts, but how do we come to terms with what they mean? Large numbers are often quoted on their own by people in authority, to try to impress the public with how big the numbers are. (A useful term might be ‘orphan’ numbers because they are not related to other numbers.) However, there is an old joke that if you ask a statistician how well they are, they will reply ‘Compared with whom?’ Likewise, numbers on their own are by and large meaningless; it is only with comparisons that we can extract a meaning. In example 1, the COVID-19 deaths are, on their own, just large numbers. However, we can employ an analogy to give them some meaning. The first number of deaths roughly equates to the same number of people at an average Premier League football club in the UK, whereas the second is closer in number to a capacity crowd at Old Trafford, home ground of Manchester United. The third is the size of an average town in the UK (e.g. Southampton). These analogies put the number of deaths into a very human perspective. However, to get a better understanding of these numbers we need more specific comparisons.

A helpful basis for comparison is knowing the approximate size of the populations to which each statistic is referring. In 2019, the population of the UK was 67 million, that of the USA was 330 million and that of the world 7.7 billion (7700 million).[6] We can then calculate the ratio of the number of deaths to the size of the population. A ratio is simply one number (numerator) divided by another (denominator). In this case, since the numerator is a subset of the denominator, we have proportions. The deaths per head of population are 0.044%, 0.023% and 0.003% for the UK, USA and the world, respectively. These percentages lead to another comparison: that between countries. The UK appears to be doing worse than the USA, which is doing worse than the rest of the world. Is this a reasonable conclusion? Cause of death is often very unreliably reported. Completion of a death certificate is often assigned to a junior doctor with little training. In an elderly patient with multiple diseases, it can be especially difficult to ascribe one main cause. So in this example we should consider how we know the person died of COVID-19. Presumably the patient was tested before they died or they had symptoms similar to COVID-19. However, testing rates have varied widely between countries and diagnosing symptoms of COVID-19 is very subjective. Thus, these numbers for death rates due to COVID-19 are not at all reliable and a reliable comparison is therefore difficult.

In contrast, deaths (from any cause) are reliably reported in the UK and the USA and probably well reported for the rest of the world. In example 2, again the numbers by themselves are not meaningful, but compared to the size of the relevant populations we can extract some meaning. A quick calculation reveals that 0.95% of the UK population dies every year, compared to 0.88% in the USA and 0.76% in the world. These numbers on their own are interesting. In the UK about 1 person in 100 dies each year. This brings the numbers down to something we can appreciate. Again, we can compare the proportions dying by country, and once more it appears that the UK is so much less healthy than the USA, and both countries are less healthy than the rest of the world. This may lead to further investigations.

In example 3, we could compare the extra sum invested in the NHS to the annual budget for the NHS, which is about £130 billion, so this extra £1 billion is less than 1% of the total. Another way to look at this is to consider that we now know there are about 67 million people in the UK, so £1 billion equates to about £15 for every person in a year, roughly the cost of five pints of beer (at current UK prices outside of London). It doesn’t sound so big now, does it?

In example 4, it is worth knowing that the UK economy was worth £8.8 trillion a year in 2016 (a trillion is 1000 billion).[7] The £350 million a week given to the EU is £18.2 billion a year, so the amount the UK sends to Europe is $\frac{18.2}{8800} \times 100 = 0.2\%$ of the UK economy. Again, it doesn't sound so big now, does it?

If we combine the information from example 5 with the worldwide death data in example 2, we would deduce that approximately 1 in 5 deaths worldwide is due to sepsis. This certainly is a large number! However, all unusual numbers should be subjected to a little scrutiny. As a quick reality check, you might start by asking yourself whether of the people you know who died recently, did 1 in 5 die of sepsis? One would expect the answer to be no. Thus, we might query whether the GBDR is right. One issue is that sepsis can be difficult to diagnose and the rate of diagnosis varies hugely from one country to another, so local experience may be misleading in that in another country sepsis might be more readily diagnosed.

When you hear a number given that you believe the presenter wants to sound big, it is always worth applying reality checks such as those described in Box 1.1. A light-hearted example has been provided in a video from the Sheffield Methods Research Institute[8] concerning a news report that stated that floods in New Zealand had caused 30,000 pigs to be washed down a river. This was then reported uncritically by other news outlets, until someone thought: '30,000? That's an enormous number, is it believable? How many pigs are likely to fall into a river at any one time?' Going back to the original broadcast, it turned out that the reporter had in fact said 30 *sows and* pigs, but owing to their New Zealand accent, this got 'misheard' and repeated uncritically to the wider media.

A further question about a large number is to ask what period of time the number refers to. By expanding or contracting the time scale, a presenter can make a number look big or small. When a large sum of money is promised, one should ask: How much does this equate to per year? In example 2 above, 634,000 deaths sounds large, whereas 0.95% sounds small. However, if one stated that approximately 1650 people die every day, or about 1 every minute, it may sound even bigger, since in our everyday experience people are not dying every minute! It is worse when reports state the 'risk of death' or 'lives saved' without stating a time period. The risk of death in the long run is one!

Definitions of the quantities discussed in this chapter are given in the Glossary. Ways of questioning numbers are given in Box 1.1.

Ratios

As we have shown, a number on its own is difficult to comprehend, but when compared to another number it can be given meaning. The simplest way to compare two numbers is to divide one by the other. A ratio is simply one number (numerator) divided by another (denominator). Ratios of continuous variables are often used to 'adjust' the numerator by the denominator. Possibly the most commonly used ratio in medicine is the Body Mass Index (BMI), which is a person's weight in kilograms divided by their height in metres squared (kg/m^2). The idea is that tall people are naturally heavier than small people because they are bigger, but that doesn't make all tall people overweight! To decide whether someone is overweight, you can't just compare weight, you need to make some allowance for height. The idea of dividing by the square of height is credited to Adolphe Quetelet (1796–1874), who observed in a cross-sectional study that weight increased as the square of height. However, it may seem simplistic to think that a simple ratio can ‘adjust’ for the denominator. For example, the BMI has received much criticism in that it doesn’t properly account for height, thus is more likely to classify short people as overweight, and also because it doesn’t account for muscle mass, which is more dense than fat. Consider that Arnold Schwarzenegger and Tom Cruise are both estimated to have a BMI over 30, which classifies them as obese![9]

Box 1.1 Things to think about to help understand numbers

Where did the number come from?

Why is this number being given and what is it supposed to show?

Can one trust the source?

Is there a useful comparator?

If there is a comparator, why was it chosen?

Is it the best one?

What period of time is the number covering?

If the number is a proportion, is the numerator relevant to the denominator?

If the number is the death rate of people with a disease, ask: How do we know that the people who died (numerator) had the disease? How do we know whether people who did not know had the disease?

What is the size of the population from which the number is coming?