APPENDIX 2: EHINZ Analytical Toolkit & Glossary

Output Estimates	Calculation	Suppression Rules (for unreliable/identifiable estimates)	Simple Definition/Explanation	Resources/References	Additional Notes/Examples
Count **a.k.a. Numerator**	*Number of individuals in a population with a condition of interest.* This is sometimes referred to as 'cases'.	**<5 in numerator** (suppressed only if the output is identifiable due to a small number for a small subgroup within a small geographic area) When dealing with small numerator counts, consider aggregating years, age groups, geographical areas or diagnostic groups. N.B. Be cautious not to aggregate subgroups that are different to each other.	A 'count' is the number from a population with a condition of interest. Counts are suppressed if it could lead to identifying individual persons.	*APHO - Technical Briefing 3, Commonly used public health statistics and their confidence intervals.pdf	
Population **a.k.a. Denominator**	*Number of individuals in a defined group.*		A 'population' is the total number in a defined group.	*APHO - Technical Briefing 3, Commonly used public health statistics and their confidence intervals.pdf	
Mean	$Mean = \frac{Sum\ of\ values\ in\ a\ sample}{Number\ of\ values\ in\ the\ sample}$	**<30 in denominator**	A 'mean' is the average number (ie, adding all data points and dividing by the number of data points).	*APHO - Technical Briefing 3, Commonly used public health statistics and their confidence intervals.pdf *MoH - NZHS Methodology Report, Analysis methods.pdf (pages 28-34)	
Proportion	$Proportion = \frac{Count}{Population}$ This is sometimes refered to as 'prevalence'.	**<5 in numerator** **or <30 in denominator** When dealing with small numerator counts, consider aggregating years, age groups, geographical areas or diagnostic groups. N.B. Be cautious not to aggregate subgroups that are different to each other.	A 'proportion' is the count divided by the population. Proportions are suppressed if the count<5 or population<30, due to unreliability of the estimate with small numbers.	*APHO - Technical Briefing 3, Commonly used public health statistics and their confidence intervals.pdf *MoH - NZHS Methodology Report, Analysis methods.pdf (pages 28-34)	
Percent	$Percent = \frac{Count}{Population} . 100$ If both the numerator and denominator are missing then the 95% confidence interval cannot be calculated.	**<5 in numerator** **or <30 in denominator** When dealing with small numerator counts, consider aggregating years, age groups, geographical areas or diagnostic groups. N.B. Be cautious not to aggregate subgroups that are different to each other.	A 'percent' is the proportion multiplied by 100. Percents are suppressed if the count<5 or population<30, due to unreliability of the estimate with small numbers.	*APHO - Technical Briefing 3, Commonly used public health statistics and their confidence intervals.pdf *MoH - NZHS Methodology Report, Analysis methods.pdf (pages 28-34)	
Crude Rate (CR)	$CR = \frac{Count}{Population} . 100{,}000$ Also applies to calculating a rate for a specific age-group (sometimes referred to as an age-specific crude rate). If the numerator uses the sum of the counts across multiple years, the denominator should use the sum of the population over the same years.	**<5 in numerator** **or <30 in denominator** When dealing with small numerator counts, consider aggregating years, age groups, geographical areas or diagnostic groups. N.B. Be cautious not to aggregate subgroups that are different to each other.	A 'crude rate' represents the proportion of the population affected by a disease. It takes into account the population size. It does not take into account varying age distributions when comparing between populations. Crude rates are suppressed if the count<5 or population<30, due to unreliability of the estimate with small numbers.	*APHO - Technical Briefing 3, Commonly used public health statistics and their confidence intervals.pdf *PHC/MoH - Standardising Rates of Disease.pdf (page 6) *MoH - NZHS Methodology Report, Analysis methods.pdf (pages 28-34)	"For some groups, where suicide numbers are very small, rates can fluctuate greatly between years and are not reliable. These rates are not suitable to be used in statistical comparisons as they can lead to misleading conclusions. For this reason rates are not calculated where a category has fewer than five suicide deaths per year." https://www.health.govt.nz/publication/suicide-facts-data-tables-19962016
Age-Standardised Rate (ASR)	$ASR = \frac{\sum P_i r_i}{\sum P_i} . constant(unit\ of\ population)$ where P_i=the population (or weight) in age group i of the standard population and r_i=the rate in age group i of the study population The calculation uses the direct standardisation method. It uses 5-year age-bands and standardises to the WHO World Standard Population. See the 'Standardising Rates of Disease' document for further details. NB: the rate itself is arbitrary, as it is standardised to an imaginary standard population.	**<20 in numerator (all ages)** **or <30 denominator in any age-band of calculation** When dealing with small numerator counts, consider aggregating years, age groups, geographical areas or diagnostic groups. N.B. Be cautious not to aggregate subgroups that are different to each other.	An 'age-standardised rate' is standardised to the World Health Organisation (WHO) population with a standard age distribution, to allow for comparisons between populations with differing age structures. Age-standardised rates are suppressed if the overall count<20, due to unreliability of the estimate with small numbers, or if any age-band of the calculation has a population<30.	*APHO - Technical Briefing 3, Commonly used public health statistics and their confidence intervals.pdf *PHC/MoH - Standardising Rates of Disease.pdf (page 11) *MoH - NZHS Methodology Report, Analysis methods.pdf (pages 28-34) *AIHW - Principles on the use of direct age-standardisation in administrative data collections.pdf *WHO - Age Standardization of Rates, a new WHO Standard (2000-2025).pdf (Table 1, page 10)	Follow principles of Box1(pg10) in AIHW document. However, replace the standard population in Principle 1 with the World Health Organisation (WHO) population.

Standardised Mortality Ratio (SMR)	$SMR = \frac{total\ observed\ deaths}{total\ expected\ deaths} . 100$ The standardised mortality ratio (SMR) is calculated using the indirect age-standardisation method. See the 'Standardising Rates of Disease' document for further details.		A 'standardised mortality ratio' can be compared to a standard.	*APHO - Technical Briefing 3, Commonly used public health statistics and their confidence intervals.pdf *PHC/MoH - Standardising Rates of Disease.pdf (page 16)	
Rate Difference (RD)	$Rate_1 - Rate_2$ Where possible, consider creating a difference ≥0 for easier interpretation (ie, using the larger group in the second rate, $Rate_2$).	If $rate_1$ or $rate_2$ meets any suppression criteria then suppress the rate difference.	A 'rate difference' is the difference between two rates. Rate-differences are suppressed if either of the rates are suppressed. A rate-difference is reported if the difference is statistically significant at the 5% level or less. (See page 3 for further details)	*Milo Schield - Common errors in forming arithmetic comparisons.pdf	"X.X percentage points more/less than" Eg, a rate difference of (60% - 50%) is "10 percentage points higher than". We don't use the words 'percent more than' as it gets confusing and can be misinterpreted. Eg, "There was a decrease in rate between 2001 (X.X per 100,000 population) and 2017 (Y.Y per 100,000 population)". We also note: "Unless otherwise stated, all differences between two values mentioned in the text are statistically significant at the 5% level or less."
Rate Ratio (RR)	$\frac{Rate_1}{Rate_2}$ Where possible, consider creating a ratio ≥1 for easier interpretation (ie, using the larger group in the denominator rate, $Rate_2$). Present all Rate Ratio output to 2 decimal places (2dp).	If $rate_1$ or $rate_2$ meets any suppression criteria then suppress the rate ratio.	A 'rate ratio' is a ratio of two rates. Rate-ratios are suppressed if either of the rates are suppressed. A rate-ratio is reported if the ratio is statistically significant at the 5% level or less. (See page 3 for further details)	*Milo Schield - Common errors in forming arithmetic comparisons.pdf	"X.XX times as much as" Eg, 3/1 is "three times as much as".

NB: Present all output to 1 decimal place (1dp), unless otherwise required (eg, for Rate Ratios).
NB: Present all rates to 100,000 population, unless otherwise required.
Key:
† see Confidence Interval (CI) in the Other Statistical Terms table below.

Links relating to Resources/References
APHO: https://fingertips.phe.org.uk/profile/guidance
MoH: https://www.health.govt.nz/nz-health-statistics/national-collections-and-surveys/surveys/new-zealand-health-survey
PHC/MoH: https://www.health.govt.nz/system/files/documents/publications/standardising-rates-disease_0.pdf
AIHW: https://www.aihw.gov.au/getmedia/95237794-4b77-4683-9f00-77c4d33e0e7c/13406.pdf.aspx?inline=true
WHO: https://www.who.int/healthinfo/paper31.pdf
Milo Schield: https://web.augsburg.edu/~schield/MiloPapers/984OfSigCompare3.pdf

Other Statistical Terms	Calculation	Suppression Rules (for unreliable/identifiable estimates)	Simple Definition/Explanation	Resources/References	Additional Notes/Examples
Sample Standard Deviation (SD)	$SD = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{(n-1)}}$	Suppress if the corresponding estimate is suppressed.	A 'standard deviation' is a measure of the spread of a sample.	*APHO - Technical Briefing 3, Commonly used public health statistics and their confidence intervals.pdf	
Standard Error (SE)	$SE = \frac{SD}{\sqrt{n}}$	Suppress if the corresponding estimate is suppressed.	A 'standard error' is a measure of precision of an estimate from it's expected value.	*APHO - Technical Briefing 3, Commonly used public health statistics and their confidence intervals.pdf	
Confidence Interval (CI)		Suppress if the corresponding estimate is suppressed.	A 'confidence interval' is a range around an estimate to quantify how precise the estimate is. The wider the confidence interval, the greater the uncertainty in the estimate. The level of confidence is an arbitrary value giving the desired probability that the interval includes the true value. We use a default of 95% confidence, unless otherwise stated. (See page 3 for further details)	*PHE (1) - Technical Guide, Confidence Intervals.pdf *APHO - Technical Briefing 3, Commonly used public health statistics and their confidence intervals.pdf	See pages 4-6 of the APHO document for more information about confidence intervals, including using confidence intervals to make comparisons.

Links relating to Resources/References
APHO: https://fingertips.phe.org.uk/profile/guidance
PHE: https://fingertips.phe.org.uk/profile/guidance

Statistical significance and p-values

Statistical tests come in a variety of forms. What they have in common is that they test hypotheses. A simple example is whether a mean from a sample is from a population with mean equal to a specified value. What is called the null hypothesis is that it does, and the alternative hypothesis is that it does not.

Sample data is not enough to make this decision with certainty. What we do instead is assess how likely the null hypothesis is given the data. This probability (based on an appropriate statistical test) is called a p-value.

A p-value can be anything between 0 and 1. If it is small enough, the null hypothesis (in our example that the mean equals a particular value) is unlikely, so we choose the alternative hypothesis (that the population mean does not equal this value). Common cut-off points are $p \leq 0.05$ and $p \leq 0.01$ (corresponding to one chance in 20 and one chance in 100 respectively) but other cut offs are also used.

Confidence intervals

When results are plotted, it is sometimes useful to indicate how accurate they are. This can be done via an extension of p-values. Suppose for example we use p=0.05, so that 1-p=0.95=95%, and then calculate where above and below the mean this p-value corresponds. These points are called the upper and lower confidence bounds and the distance between them the 95% confidence interval. They are often shown as a bar that extends either side of a statistic. For a mean, based on a normal distribution, these are plus and minus two (or more accurately 1.96) times the standard error either side of the mean, but for some statistics such as ratios their difference from the sample statistic is not based on 1.96 and is not symmetric.

Why a small difference may be statistically significant when a larger one is not

When we do statistical tests, they can only be based on the information we have, not in an ideal world where everything is known. Indeed, if we knew the true population value of statistic (since it would be known without error) there would be no need for a statistical test. When we compare two groups which are of different sizes (eg, different age groups or ethnic groups) with an overall value, the statistical test depends not just on how different the sample information for each group is from this chosen overall value but also on how big the two groups are. If one group is smaller, its sample estimate (eg, its mean) will be known less accurately, so that even if it differs more from the overall value it is possible that it has a larger p-value (ie, it will be less significantly different) than for the larger group. It comes back to what information we have and how accurate it is, not simply how big differences are between groups' sample values and the overall value.

This issue is a fundamental one for statistical hypothesis testing. Statistical tests depend not just on sample values but on how accurately they are known. So simply looking at the size of sample differences and comparing them on that basis alone is not sufficient.

The same concept applies when testing means (or other statistics such as rates or ratios) for groups against an overall value which is itself an estimate rather than a fixed value, or when there are more than two groups and each group is tested against each other.

Population and sample sizes

Sometimes in statistical tests, population sizes are used instead of sample sizes. In one sense, since we have a population size we do not need a statistical test. But often we are not interested just in this population but in some larger conceptual population. For example, we may have the population size for a particular year, but we are also interested in what may happen in other years, so the conceptual population is larger. This conceptual population is often called a super-population, and often in statistical testing the population size is then treated as a "sample size" and the super-population size if known as a "population size".

Reference

Šidák, Z. (1967) Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. Journal of the American Statistical Association; 62: 626-33.