Formula for sample size estimation

If $\bar{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error will not exceed a specified amount e when the sample size is

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2.$$

Where e, the error, is the half-width of the confidence interval.
You should round up n to the next integer value

Very useful for sample size calculations

Tolerance Intervals

- So far we have learned how we can obtain a confidence interval for the population mean for large samples (using CLT), and for small samples where the observations are normally distributed or can be transformed to normality (using the t distribution). We have also seen how to use bootstrapping to obtain a confidence interval for the mean or other parameters (eg the median).

- Irrespective of the method used, the confidence interval has the same interpretation.

What confidence intervals do

- From the Celtic study we have calculated a 95% CI for the **mean improvement in VO2 max** of youth football players who participate in the training intervention of (*4.0, 6.2) mL/kg min*
- Does this mean that each player will have an improvement in this range?
- In other words, is it meaningful to compare an individual against the mean of the population ?

Tolerance intervals

- A reference range, called a **tolerance interval**, indicating where some proportion of the population values lie can be more relevant if we are interested in a question such as how much is an **individual** likely to improve in VO2 max after the training intervention.
- The tolerance interval is estimated from the sample for a specified level of confidence - typically 95% confidence is used here also.
- It may be of interest is to find an interval that captures 95% of the population values with 95% confidence.
- This is what the **tolerance interval** does.

Tolerance Interval formula (population distribution is Normal)

- A tolerance interval for capturing at least $100(1-\gamma)\%$ of the values in a Normal distribution with confidence level $100(1-\alpha)\%$ is

$$x \ \underline{+}\ ks$$

- where *k* is a tolerance interval factor found in a Tolerance Interval table.
- The TI table gives the sample size in the left hand column, and for each sample size gives the value of k needed for various combinations of$100(1-\gamma)\%$ (the population proportion required) and $100(1-\alpha)\%$ (the confidence level).

Tolerance Interval Table – Normal distribution

Values of k for Two-Sided Intervals									
Confidence Level	0.90			0.95			0.99		
Percent Coverage	0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
2	15.978	18.800	24.167	32.019	37.674	48.430	160.193	188.491	242.300
3	5.847	6.919	8.974	8.380	9.916	12.861	18.930	22.401	29.055
4	4.166	4.943	6.440	5.369	6.370	8.299	9.398	11.150	14.527
5	3.949	4.152	5.423	4.275	5.079	6.634	6.612	7.855	10.260
6	3.131	3.723	4.870	3.712	4.414	5.775	5.337	6.345	8.301
7	2.902	3.452	4.521	3.369	4.007	5.248	4.613	5.488	7.187
8	2.743	3.264	4.278	3.136	3.732	4.891	4.147	4.936	6.468
9	2.626	3.125	4.098	2.967	3.532	4.631	3.822	4.550	5.966
10	2.535	3.018	3.959	2.839	3.379	4.433	3.582	4.265	5.594
11	2.463	2.933	3.849	2.737	3.259	4.277	3.397	4.045	5.308
12	2.404	2.863	3.758	2.655	3.162	4.150	3.250	3.870	5.079
13	2.355	2.805	3.682	2.587	3.081	4.044	3.130	3.727	4.893
14	2.314	2.756	3.618	2.529	3.012	3.955	3.029	3.608	4.737
15	2.278	2.713	3.562	2.480	2.954	3.878	2.945	3.507	4.605
16	2.246	2.676	3.514	2.437	2.903	3.812	2.872	3.421	4.492
17	2.219	2.643	3.471	2.400	2.858	3.754	2.808	3.345	4.393
18	2.194	2.614	3.433	2.366	2.819	3.702	2.753	3.279	4.307
19	2.172	2.588	3.399	2.337	2.784	3.656	2.703	3.221	4.230
20	2.152	2.564	3.368	2.310	2.752	3.615	2.659	3.168	4.161
21	2.135	2.543	3.340	2.286	2.723	3.577	2.620	3.121	4.100
22	2.118	2.524	3.315	2.264	2.697	3.543	2.584	3.078	4.044
23	2.103	2.506	3.292	2.244	2.673	3.512	2.551	3.040	3.993
24	2.089	2.489	3.270	2.225	2.651	3.483	2.522	3.004	3.947
25	2.077	2.474	3.251	2.208	2.631	3.457	2.494	2.972	3.904
30	2.025	2.413	3.170	2.140	2.529	3.350	2.385	2.841	3.733
40	1.959	2.334	3.066	2.052	2.445	3.213	2.247	2.677	3.518
50	1.916	2.284	3.001	1.996	2.379	3.126	2.162	2.576	3.385
60	1.887	2.248	2.955	1.958	2.333	3.066	2.103	2.506	3.293
70	1.865	2.222	2.920	1.929	2.299	3.021	2.060	2.454	3.225
80	1.848	2.202	2.894	1.907	2.272	2.986	2.026	2.414	3.173
90	1.834	2.185	2.872	1.889	2.251	2.958	1.999	2.382	3.130
100	1.822	2.172	2.854	1.874	2.233	2.934	1.977	2.355	3.096

Tolerance interval for VO2 max improvement

Suppose we want to use the Celtic study to make an interval estimate where we are 95% confident that 95% of the VO2 max improvements will lie.

From the tolerance interval table, the tolerance factor k for n = 18, required proportion $100(1-\gamma)\%$ 95%, and 95% confidence is k = 2.819.

- The 95% 95% tolerance interval then is

$$5.11 \pm 2.819 * 2.26 = (-1.3, 11.5)$$

Interpretation of Tolerance Interval

We can be 95% confident that at least **95% of youth players who have the training intervention** will improve between -1.3 and 11.5 ml/Kg min in VO2 max.

- The tolerance interval covers a wider range of values than the confidence interval for the mean.
- The width of confidence interval for the mean depends on the **sampling error of the sample mean**
- If we sample all individuals in the population the confidence interval has width zero!
- The tolerance interval width depends on both **random variation of the individual values** and the **sampling error**
- If the sample size increases, the tolerance interval will converge to the range of the values in the population within which the required proportion $100(1-\gamma)\%$ lies.

Tolerance intervals in R

The **tolerance** package in R provides functions to calculate tolerance intervals for different distributions and under different scenarios.

The **`normtol.int()`** function can be used to provide tolerance intervals for data distributed according to either a Normal distribution or a log-Normal distribution.

normtol.int(x, alpha = 0.05, P = 0.99, side = 2, log.norm = FALSE)

- `x` represents a vector of data which is distributed according to either a Normal distribution or a log-Normal distribution.
- `alpha` represents the level chosen such that `1 - alpha` is the confidence level,
- -`P` represents the proportion of the population to be covered by this tolerance interval, P = $(1-\gamma)$
- - `side` indicates whether a 1-sided or 2-sided tolerance interval should be generated. The default is `side = 1`. For the purpose of this course we use `side = 2`.
- - `log.norm = TRUE` will be used when the logarithm transformation of data is Normally distributed.

Tolerance interval for VO2 max improvement using R

```
>
> normtol.int(train.df$Improvement, alpha = 0.05, P = 0.95, side = 2)
  alpha    P    x.bar 2-sided.lower 2-sided.upper
1  0.05 0.95 5.111111     -1.283792      11.50601
>
```