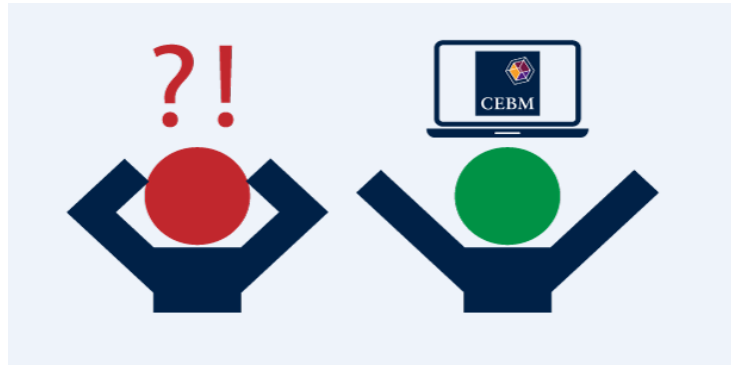


## Tip for data extraction for meta-analysis – 26



### What if you are missing a standard deviation and only a similar summary statistic is given?

Kathy Taylor

[Previously](#), I highlighted a list of ways where, when extracting data for meta-analysis of continuous outcomes, you might find that a summary statistic that you want is missing. In my last post [I gave the 3<sup>rd</sup> way - a similar summary statistic is reported, but it's not the statistical measure that I want](#) and I focused on missing means. In this post I'll show you what you can do with missing standard deviations (SDs).

Instead of the SD, another measure of dispersion may be reported, either the standard error (SE), confidence interval (CI), interquartile range (IQR) or range. The SD describes how measurements of participants naturally differ (which is saying something about the population) whilst the SE describes how accurately the mean has been estimated (which is saying something about a study). Sometimes it's not what clear if the reported statistic is the SE or the SD and so comparing its value with the established SEs or SDs of other studies may help you decide.

The [Cochrane Handbook \(6.5.2.2.\)](#) divides the equations for calculating SDs into those for group means (when you want the SD of a mean value for the intervention group or the control group) and difference in means (when you want the SD of a difference in means between the intervention and control groups). In this post I deal with SDs of group means and I will look at SDs of difference in means and other effect measures in a future post.

#### *Calculating SDs from SEs:*

Obtaining SDs from SEs is very simple

$$SD = SE\sqrt{n}$$

### *Calculating SDs from confidence intervals:*

A 95% confidence interval is expressed in terms of the SE and gives the range in which we are 95% sure that the sample mean lies. For data that is normally distributed, the confidence interval will be symmetric about the mean and therefore,

$$SE = \frac{(\text{upper CI} - \text{lower CI})}{3.92}$$
$$SD = \frac{(\text{upper CI} - \text{lower CI})}{3.92} \sqrt{n}$$

For a 90% confidence interval, divide by 3.29, and for a 99% confidence interval, divide by 5.15. These divisors are derived from the standard normal distribution. If the sample size is small (<60 in each group), the divisors should be replaced by slightly larger numbers, derived from the t-distribution. Tables for these two distributions are given at the end of this post.

### *Calculating SDs from IQRs:*

The [Cochrane Handbook](#) states that for normally distributed data, you can estimate

$$SD = \frac{IQR}{1.35}$$

### *Calculating SDs from other summary statistics:*

There are a number of ways of calculating the SD from the range but they are not generally recommended by [Cochrane Handbook](#) because the range is so unstable, as it is determined by extreme values rather than providing an average measure of variation.

A common approach is to estimate

$$SD = \frac{\text{range}}{4}$$

[Walter and Yao](#) provide a table of conversion factors (f) according to the sample size to estimate

$$SD = f \times \text{range}$$

Their table suggest that the common formula only applies to a sample size of around size 25 (f=0.254).

Other methods estimate the SD by equations of several other statistics. These equations have been evaluated by simulation but not empirically so the [Cochrane Handbook \(section 6.5.2.6\)](#) do not recommend them “as a general rule” but these estimates could still be used and the studies removed in a sensitivity analysis.

[Hozo et al](#) provide an estimate of the SD using the range with the median and sample size

$$SD = \sqrt{\frac{n+1}{48n(n-1)^2} ((n^2+3)(\min - 2\text{median} + \max)^2 + 4n^2(\max - \min)^2)}$$

which they simplify for large n to

$$SD = \sqrt{\frac{1}{12} \left( \frac{(\min - 2\text{median} + \max)^2}{4} + (\max - \min)^2 \right)}$$

[Bland](#) provides an estimate based on the range and interquartile range with the mean and sample size:

$$SD = \sqrt{\frac{\frac{FUNCT}{16} - n \times \text{mean}^2}{n-1}}$$

Where

$$\begin{aligned} FUNCT = & 2(n+3)(q_1^2 + \text{median}^2 + q_3^2) \\ & + 2(n-5)(\min \times q_1 + \text{median} \times q_1 + \text{median} \times q_3 + \max \times q_3) \\ & + (n+11)(\min^2 + \max^2) \end{aligned}$$

[Wan et al](#) estimate the SD from the range with the median and sample size:

$$SD \approx \frac{\text{range}}{2\Phi^{-1}\left(\frac{n-0.375}{n+0.25}\right)}$$

They estimate the SD from the range, interquartile range, median and sample size,

$$SD \approx \frac{\text{range}}{4\Phi^{-1}\left(\frac{n-0.375}{n+0.25}\right)} + \frac{q_3 - q_1}{4\Phi^{-1}\left(\frac{0.75n-0.125}{n+0.25}\right)}$$

and from the interquartile range and sample size (for large sample sizes)

$$SD \approx \frac{q_3 - q_1}{2\Phi^{-1}\left(\frac{0.75n-0.125}{n+0.25}\right)}$$

Where

$\Phi^{-1}(z)$  is the inverse function of  $\Phi(z)$  (the cumulative distribution function of the standard normal distribution).  $\Phi^{-1}(z)$  is also the upper zth percentile of the standard normal distribution. It can be calculated using the R software command 'qnorm(z)'.

## Examples of studies with missing data

Let me show you some examples from studies of people with diabetes which were included in systematic reviews carried out by our group.

A study by [Chaisson et al 2001](#) reported the effect of metformin on change from baseline of HbA1c in terms of mean and SE.

For the intervention group

$$SD = 0.12\sqrt{81} = 1.08\%$$

For the control group

$$SD = 0.12\sqrt{82} = 1.09\%$$

[Kemal et al](#) reported the effects of rosiglitazone on plasma glucose and other laboratory variables at 6 months in terms of median and range.

Three studies from one review where we extracted data on the effects of renin-angiotensin-aldosterone system inhibitors on albumin excretion rates were [Tan et al](#) who reported the effects of losartan at 6 months in terms of the median and interquartile range (IQR). [Bojestig et al](#) reported the effects of ramipril at 2 years in terms of median and range, and [Tong et al](#) reported the effects of fosinopril, also at 2 years in terms of median and range. Table shows the SD calculations using the different equations that I have shown above. Albumin excretion is measures in  $\mu\text{g}/\text{min}$  for all studies. For Tong et al, I converted the data from mg/24 hours, using the conversion factor that I showed [previously \(no.5\)](#).

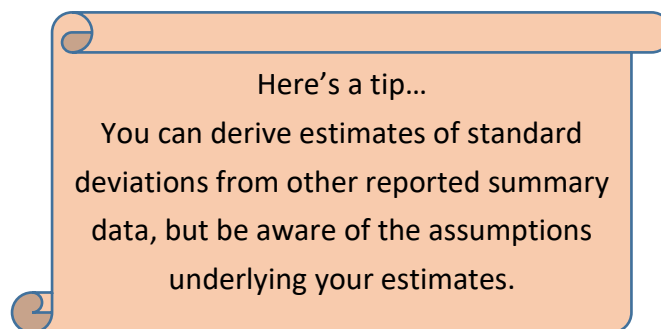
Table. Estimating standard deviations

Study	Tan et al		Bojestig et al			Tong et al		Kemal et al	
DATA									
Statistic	Intervention	Control	Intervention - L	Intervention - H	Control	Intervention	Control	Intervention	Control
n	40	40	16	17	18	18	20	11	17
median	79	55	81	94	96	894	243	2.71	2.64
lower IQR						103	107		
upper IQR						3318	1836		
IQR	101	58				3215	1729		
min			10	23	48				
max			1450	1112	308				
Range/4			1440	1089	260			2.38	1.55
f			0.283	0.279	0.275			0.315	0.279
SD ESTIMATIONS									
Equation	Intervention	Control	Intervention - L	Intervention - H	Control	Intervention	Control	Intervention	Control
Common			360.00	272.25	65.00			0.60	0.39
Walter & Yao			407.52	303.83	71.50			0.75	0.43
Wan et al 1			407.05	272.25	65.00			0.75	0.43
Wan et al 2	77.66	44.60				2586.33	1379.48		
Cochrane	74.81	42.96				2381.17	1280.86		

Common approach – range/4; Cochrane Handbook – IQR/1.35

For the data from Tan et al, the equations of Wan et al and Cochrane Handbook produce similar results, which suggests that the distribution of the data were not highly skewed as the latter equation is based on assumption that the data are normally distributed. A similar point could be made for Tong et al. For the data reported by Kemal et al, the equations of Wan et al and Walter and Yao produced identical results to 2 decimal places, but the simple common approach underestimated the SDs. Applying the equations to the data of Bojestic et al shows how wide ranges can produce unstable results.

Another strategy which I will cover in my next post is dealing with missing SDs by imputation. Which SD should you use? Take an average, use the lowest value or highest value, or try them all? I will address these questions in a future post on sensitivity analysis.



In my next post, I'll focus on some other examples of the [4<sup>th</sup> way](#) of how a summary statistic that you want may be missing for some cases: **neither the summary statistic you want, nor a similar statistic are reported.**

Where did the equations come from?

(You can skip this if you are only interested in carrying out the calculations)

*Calculating SDs from SEs:*

The standard error of the mean (SEM, which is often abbreviated to SE) is the standard deviation of the means of multiple samples:

$$SE = \frac{\sigma}{\sqrt{n}}$$

Where

n= sample size

$\sigma$  = population standard deviation

The SE can be estimated from a single sample using the observed sample standard deviation, s:

$$SE \approx \frac{s}{\sqrt{n}}$$

Let  $x_1, x_2, x_3, \dots, x_n$  be n independent observations from a population with mean  $\mu$  and standard deviation  $\sigma$  (and variance  $\sigma^2$ )

$$T = x_1 + x_2 + \dots + x_n$$

$$Var(T) = Var(x_1 + x_2 + \dots + x_n) = n\sigma^2$$

$$\bar{x} = \frac{T}{n}$$

$$Var(\bar{x}) = Var\left(\frac{T}{n}\right) = \frac{1}{n^2}Var(T) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

This used the result

$$Var(aX) = a^2Var(X)$$

which comes from

$$Var(X) = E((X - \mu)^2) \quad \text{where } \mu = E(X)$$

$$Var(X) = E(X^2) - 2E(X)\mu + \mu^2$$

$$Var(X) = E(X^2) - 2\mu^2 + \mu^2$$

$$Var(X) = E(X^2) - \mu^2$$

$$Var(X) = E(X^2) - (E(X))^2$$

Therefore,

$$Var(aX) = E((aX)^2) - (E(aX))^2$$

$$Var(aX) = a^2E(X^2) - a^2(E(X))^2 = a^2Var(X)$$

*Returning to*

$$Var(\bar{x}) = \frac{\sigma^2}{n}$$

$$SD \text{ of } \bar{x} = SEM = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

*Rearranging*

$$SD = SE\sqrt{n}$$

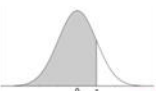
*Calculating SDs from confidence intervals:*

If we call the upper and lower limits of the 95% confidence interval upperCI and lowerCI. A symmetric confidence interval means that

$$upperCI = mean + 1.96SE$$

$$lowerCI = mean - 1.96SE$$

1.96 is the Z value taken from the standard normal distribution table with the area in each tail of  $(1-0.95)/2=0.025$  and therefore, using the one-sided table (Figure 1, red), the shaded area is  $1-0.025=0.975$



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.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

Figure 1. Standard normal distribution table (p=0.95,0.975,0.995)

[As shown before](#), rearranging the equations for upperCI and lowerCI

$$(2 \times 1.96)SE = upperCI - lowerCI$$

Rearranging,

$$SE = \frac{(upper\ CI - lower\ CI)}{3.92}$$

Similarly, for a 90% confidence interval, the area in each tail is  $(1-0.90)/2=0.05$  and the shaded area corresponding to a one-sided standard normal distribution table is  $(1-0.05)=0.95$ . The corresponding z value is 1.645 (Figure 1, green).

$2 \times 1.645 = 3.29$  and therefore,

$$SE = \frac{(upper\ CI - lower\ CI)}{3.29}$$

For a 99% confidence interval, the area in each tail is  $(1-0.99)/2=0.005$  and the shaded area corresponding to a one-sided standard normal distribution table is  $(1-0.005)=0.995$ . The corresponding z value is 2.575 (Figure 1, blue).

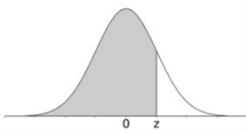
$2 \times 2.575 = 5.15$  and therefore,

$$SE = \frac{(upper\ CI - lower\ CI)}{5.15}$$

### Calculating SDs from IQRs:

From a standard normal distribution table (Figure 2), the Z value for shaded area 0.75 (upper quartile) is approximately 0.67. The upper quartile is 0.67 SDs from the mean so

$$IQR = 2 \times 0.67 \times SD \approx 1.35 SD$$



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.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

Figure 2. Standard normal distribution table (p=0.75)

### Calculating SDs from other summary statistics:

Walter and Yao provide information about the sources of their table of conversion factors. Estimates of Hozo et al, Bland and Wan et al all provide detailed derivations of their equations in their papers. Wan also provide an online spreadsheet to calculate and compare their estimates. The common estimate of the SD as ¼ of the range comes from the fact that in normally distributed data, approximately 95% of values lie between 2 standard deviations either side of the mean (Figures 3). The shaded area in the one sided standard normal table is 1-0.0228=0.9972 (Figure 4).



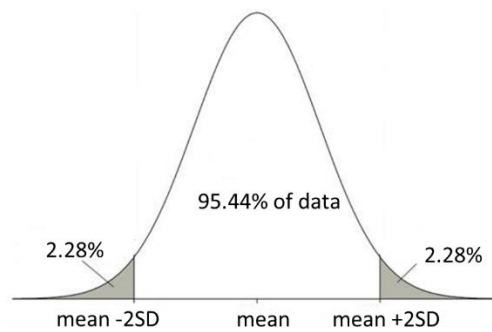
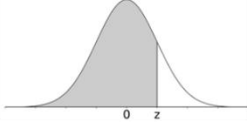


Figure 3. Probability of being within  $\pm 2SD$  of the mean for data normally distributed



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.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

Figure 4. Standard normal distribution table ( $p=0.9772$ )

So ignoring the 4.56% in the tails, the range is estimated as

$$range = 4SD$$

The estimate of the SD then follows

$$SD = \frac{range}{4}$$

Dr Kathy Taylor teaches data extraction in [Meta-analysis](#). This is a short course that is also available as part of our [MSc in Evidence-Based Health Care](#), [MSc in EBHC Medical Statistics](#), and [MSc in EBHC Systematic Reviews](#).

Follow updates on this blog, related news, and to find out about other examples of statistics being made more broadly accessible on Twitter [@dataextips](#)