

(FREQUENTIST) INFERENCE for (PARAMETER) μ – the mean of an (effectively) infinite-size universe of Y values – based on the n values, y_1, \dots, y_n in an SRS from that universe/population

‘*Certain conditions apply*’¹

Point-estimate of μ : $\hat{\mu} = \bar{y}$

(Symmetric) Confidence Interval CI for μ : $\bar{y} \pm ME$,

where the Margin of Error (ME) is a

- z -multiple of SE², if n is ‘large’ AND

the Y values in the universe have a ‘Normal’ (Gaussian) distribution or, if not, n is large enough so that the Central Limit Theorem guarantees that the sampling distribution of possible \bar{y} ’s of size n from this universe is well enough approximated by a Gaussian distribution

- t -multiple of SE if ‘small’ n AND

the Y values in the universe have a ‘Normal’ (Gaussian) distribution

Test Statistic in relation to the Null Hypothesis $\mu = \mu_0$:

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \quad \text{or} \quad t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

The ‘ z ’ and ‘ t ’ distributions are the respective (conceptual) sampling distributions one *would* get if one

- took samples (of size n) from a Normal(μ, σ) distribution
- calculated the quantity $z = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}}$ or $t = \frac{\bar{y} - \mu}{s/\sqrt{n}}$ from each sample
- compiled a histogram of the results

‘Student’ ’s ‘curve’ for sampling distribution of $t = \frac{\bar{y} - \mu}{s/\sqrt{n}}$

- Is symmetric around 0 (just like z)

¹ B&M stress that the **first** of their conditions ‘we can regard our data as a simple random sample (SRS) from the population’ as *very important*;

the **second**, ‘Observations from the population have a Normal distribution with unknown mean parameter μ and unknown standard deviation parameter σ ’ less so: ‘*In practice, inference procedures can accommodate some deviations from the Normality condition when the sample is large enough.*’

²B&M distinguish the SD of a statistic from its SE. They switch on p. 412: ‘When the standard deviation of a statistic is estimated from data, the result is called the standard error (SE) of the statistic. The SE of the sample mean, \bar{y} is SEM = s/\sqrt{n} .’

- Has a shape like that of the Z distribution, but with a SD slightly larger than unity i.e. slightly flatter and heavier-tailed; $SD(t) = \sqrt{(n-1)/(n-3)}$. See Fig 17.1 in B&M.
- Shape becomes indistinguishable from Z distribution as $n \rightarrow \infty$ (in fact as n goes much beyond 30)
- Instead of $\pm 1.96 \times SE$ for 95% confidence (or to use as the ‘critical value’ in a null-hypothesis test), we need these multiples (or critical values):

n	‘degrees of freedom’	Multiple	from R
2	1	12.71	qt(0.975, 1)
3	2	4.30	qt(0.975, 2)
4	3	3.18	qt(0.975, 3)
11	10	2.23	qt(0.975, 10)
21	20	2.09	qt(0.975, 20)
31	30	2.04	qt(0.975, 30)
121	120	1.98	qt(0.975, 120)
∞	∞	1.96	qt(0.975, Inf)

The **width** of the t distribution is a function of how many ‘independent’ evaluations of σ one has, when **using the sample s , to estimate σ** .

σ is a measure of how far each Y in the population deviates from μ . But since we don’t know μ , we have to use the deviations of each of the n sample y ’s from the ‘best bet’ for μ , namely from the sample mean \bar{y} .

$n = 2$ is the smallest n for estimating σ : s involves the 2 deviations, $y_1 - \bar{y}$ and $y_2 - \bar{y}$. These 2 (equal size, but opposite sign) deviations add to 0, so in reality we have only one ‘independent’ evaluation of σ .

If $n = 3$, s involves 3 deviations, $y_1 - \bar{y}$, $y_2 - \bar{y}$ and $y_3 - \bar{y}$. The 3 deviations (of unequal size, with 2 of one sign balancing 1 of the other) add to 0, so we have only 2 ‘independent’ evaluations of σ : 1 combination of the 3 y ’s is used to estimate μ and the other 2 (i.e., $n - 1$) combinations measure how far the y ’s deviate from the estimate of the ‘centre.’



We refer to the ‘number of *independent* evaluations of variation’ [$(n - 1)$ here] as the **degrees of freedom**. B&M (p. 414) say students ‘often wonder’ about its meaning but they only give a minimal explanation, saying that it ‘defines the shape of a t distribution’ (and thus, a sampling distribution, and that the ‘all you need to know is that there are many t distributions, and you must specify, through the degrees of freedom, which t distribution is relevant for your computations.’

When (later) we have to estimate variation from a fitted line, we will lose a few more ‘degrees of freedom’ since it takes 2 (or more) combinations of the y ’s to fit the line from which the deviations are measured.

(Possibly-NonSymmetric) Confidence Interval CI for μ

- Bootstrap the \bar{y} !

Application: How fast is your reaction time? See related material here:
<http://www.biostat.mcgill.ca/hanley/bios601/Surveys/index.html#ReactionTimes> –
<https://www.humanbenchmark.com/dashboard> - or - <https://faculty.washington.edu/chudler/java/redgreen.html>

RED LIGHT - GREEN LIGHT Reaction Time Test			
Instructions:			
1. Click the large button on the right to begin. 2. Wait for the stoplight to turn green. 3. When the stoplight turns green, click the large button quickly! 4. Click the large button again to continue to the next test.			
Test Number	Reaction Time	The stoplight to watch.	The button to click.
1	<input type="text" value="0.325"/>		
2	<input type="text" value="0.327"/>		
3	<input type="text" value="0.357"/>		
4	<input type="text" value="0.299"/>		
5	<input type="text" value="0.378"/>		
AVG.	<input type="text" value="0.3372"/>		
<input type="button" value="Start Over"/>			

```

reaction.times = c(325,327,357,299,378)/1000
( summary(reaction.times) )
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2990  0.3250  0.3270  0.3372  0.3570  0.3780

( round( sd(reaction.times), 3) ) [1] 0.031
( n = length(reaction.times) ) [1] 5

## FROM SCRATCH (not quite 'by hand', since R does some lifting!)

( SEM = sd(reaction.times)/sqrt(n) ) [1] 0.01372734
( round(SEM,3) ) [1] 0.014
( multiple.for.95pct = qt(0.975,n-1) ) [1] 2.776445
( multiple.for.95pct.if.n.of.100 = qt(0.975,100-1) ) [1] 1.984217

```

```

( round( mean(reaction.times) +
  c(-multiple.for.95pct,0,multiple.for.95pct) * SEM,3) )
[1] 0.299 0.337 0.375

## FITTING mu using mother of all REGRESSION MODELS: E[y] = mu * 1

## ('intercept-only', '1-constant', or 'no x' linear model (lm), Least Squares fit)

fitted.model = lm( reaction.times ~ 1)
summary( fitted.model )
Residuals:
      1      2      3      4      5
-0.0122 -0.0102  0.0198 -0.0382  0.0408

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.33720    0.01373   24.56 1.63e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0307 on 4 degrees of freedom

library(MASS) ; round( confint(fitted.model), 3 )

              2.5 % 97.5 %
(Intercept) 0.299  0.375

# Interestingly, regression function 'lm' does not give CI's automatically
# It does give p-values, even if H_0 is silly

1- pt(24.56,n-1) # is the probability ABOVE 24.56 [1] 8.154983e-06
pt(-24.56,n-1) # is the probability BELOW -24.56 [1] 8.154983e-06

2*(1-pt(abs(24.56),n-1)) [1] 1.630997e-05 # is the probability
## of a more extreme result in EITHER direction (but not that relevant
## here, since a negative mu is even more silly than a zero mu)

## THE WAY MANY SOFTWARE COURSES (STILL) TEACH IT
## specialized R function -- but to get CI, to have to ask for the test !!
t.test(reaction.times)

One Sample t-test , data: reaction.times
t = 24.564, df = 4, p-value = 1.63e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.2990868 0.3753132
sample estimates: mean of x    0.3372

## FROM mosaic package ??

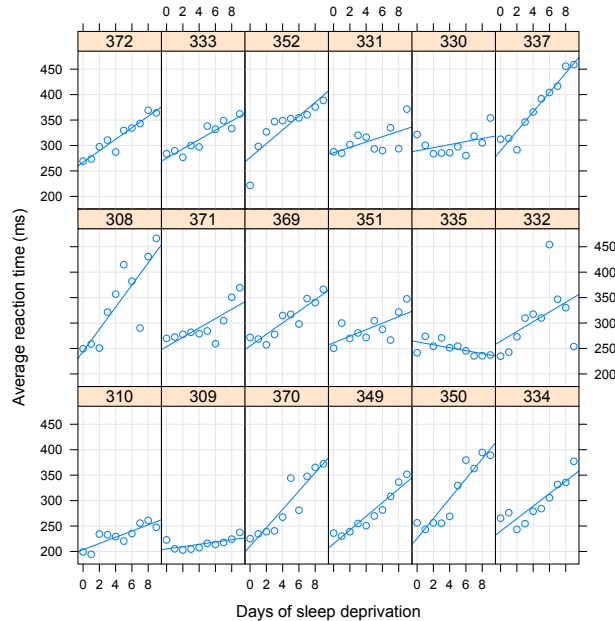
## Bootstrap ??

```

Application: Effects of sleep deprivation on reaction time

The orientational material below is from the `sleepstudy` data reanalyzed in Ch. 3 of the excellent (online) book ‘`lme4: Mixed-effects modeling with R`,’ dated June 25 2010, by Douglas M. Bates. The data are included in the `lme4` package – and were used again in the 2017 Epidemiology (teaching) article by Weichenthal, Baumgartner and Hanley.

Belenky et al. [2003] report on a study of the effects of sleep deprivation on reaction time for a number of subjects chosen from a population of long-distance truck drivers. These subjects were divided into groups that were allowed only a limited amount of sleep each night. We consider here the group of 18 subjects who were restricted to three hours of sleep per night for the first ten days of the trial. Each subject’s reaction time was measured several times on each day of the trial.



Average reaction time versus number of days of sleep deprivation by subject for the `sleepstudy` data. Each subject’s data are shown in a separate panel, along with a simple linear regression line fit to the data in that panel. The panels are ordered, from left to right along rows starting at the bottom row, by increasing intercept of these per-subject linear regression lines. The subject number is given in the strip above the panel.

Data Analysis: Each panel yields a fitted slope, the fitted prolongation of the reaction time per day of reduced sleep. We have these 18 ‘slopes’, which we denote by b_1 to b_{18} , and store in an R vector `IndividuallyFittedSlopes`

```
> round(IndividuallyFittedSlopes,1)
 [1] 21.8  2.3  6.1  3.0  5.3  9.6  9.1 12.3 -2.9 19.0 13.5 19.5  6.4 13.6 11.3
[16] 18.1  9.2 11.3 ----- 17 are positive and 1 is negative.
```

```
summary(IndividuallyFittedSlopes)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.900  6.175 10.450 10.472 13.575 21.800
sd(IndividuallyFittedSlopes) [1] 6.564325
```

FROM SCRATCH, by R

```
qt(c(0.025,0.975),17) [1] -2.109816  2.109816
round(mean(IndividuallyFittedSlopes) +
      qt(c(0.025,0.975),17) * sd(IndividuallyFittedSlopes)/sqrt(18),1)
 [1]  7.2 13.7
```

FROM SCRATCH, by ‘hand’: The mean of these 18 slopes is $\bar{b} = 10.5$ milliseconds longer per day, and their SD is $s = 6.6$ ms/day. A 95% CI for the mean delay per day is thus $\bar{b} \pm t_{17} [= 2.11] \times 6.6 / \sqrt{18} = 10.5 \pm 3.3$ ms. per day. (Allowing for the above rounding) this agrees with the `t.test` and with the fitted ‘linear model with no x’ from R:

```
t.test(IndividuallyFittedSlopes)
```

```
One Sample t-test
data: IndividuallyFittedSlopes
t = 6.7684, df = 17, p-value = 3.283e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
```

```
 7.2 13.7 [4 decimal places removed by JH]
sample estimates:
mean of x 10.5
```

```
> fit = lm(IndividuallyFittedSlopes ~ 1)
> summary(fit)
```

Coefficients:

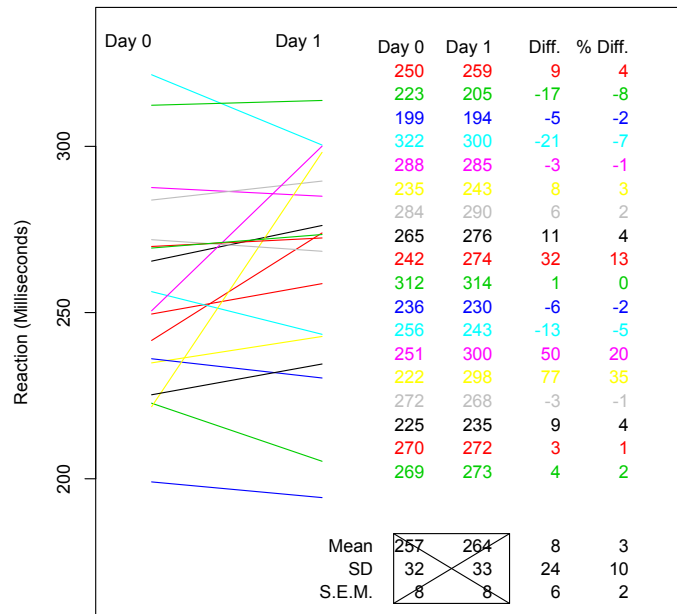
```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.472      1.547     6.768 3.28e-06 ***
```

Residual standard error: 6.564 on 17 degrees of freedom

```
> round( confint(fit), 1 )           2.5 % 97.5 %
              (Intercept)  7.2  13.7
```

There is strong evidence (see p-value) that the mean slope in the ‘population’ these 18 subjects are drawn from is non-zero.

What if just had the reaction times on day 0 and day 0?



Instead of 18 slopes across 10 days, we have 18 simple differences; they are still *slopes*, but, measured between just 2 adjacent days, are much ‘noisier.’

As in the B&M e.g., Table 17.1, p. 425, we have 18 ‘within-subject’ differences, so (just as with the 18 slopes over 10 days) it is a *single* sample of 18 differences, and so one proceeds just as with the example on the previous page. The following analyses of the *only relevant* data do not provide evidence against H_0 and the 95% CI for the mean difference includes zero.

```
IndividualDiffs.Day1.minus.Day0 = c(
9, -17, -5, -21, -3, 8, 6, 11, 32, 1, -6, -13, 50, 77, -3, 9, 3, 4)
```

```
> t.test(IndividualDiffs.Day1.minus.Day0)
```

```
One Sample t-test data: IndividualDiffs.Day1.minus.Day0
t = 1.3997, df = 17, p-value = 0.1796
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: -4.0 19.8
sample estimates: mean of x 7.9
```

Whereas the mean of 18 differences between 2 conditions is *arithmetically equal* to the difference of the 2 means of 18, the SE of the mean difference is not the same as the SE of the difference of two *independent* means.³

³In general, $\text{Var}(\bar{y}_1 - \bar{y}_0) = \text{Var}(\bar{y}_1) + \text{Var}(\bar{y}_0) - 2 \times \text{Covar}(\bar{y}_1, \bar{y}_0)$. Yet, many

Unsure about ‘Normality’ requirement for ‘t’ procedures?

```
library(mosaic)
bootstrap <- do(1000) * mean(
  resample(IndividualDiffs.Day1.minus.Day0) )
summary(bootstrap$mean)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
-5.7 4.0 7.6 7.7 11.1 25.4
confint(bootstrap) [extra -- silly-- decimal places removed]
 name lower upper level method estimate
1 mean -2.6 19.2 0.95 percentile 7.9
[COMMENT: CI not very different from model-based one]
```

HISTORICAL NOTE: These data have the same structure as those in the first illustrative example⁴ by ‘Student’ in his 1908 paper – where he derived the fore-runner of today’s t distribution: For more, see <http://www.epi.mcgill.ca/hanley/Student/> The responses of interest were: D: Difference in hours of sleep when using Drug rather than Placebo, so the Parameter of Interest is: μ_D . He used data (from an article by US doctors⁵) on $n = 10$ patients. The two inferential items are a CI for μ_D , and (the one ‘Student’ addressed) the p-value in relation to the $H_0 : \mu_D = 0$. He was not clear about whether H_{alt} was $\mu_D \neq 0$ or $\mu_D > 0$. Nor would he have been expected to be – the hypothesis-testing ‘lingo’ had not yet been formalized. Student merely calculate a variant on the p-value – which he reported as an odds. See details on page 2 of these 607 Notes from 2001: http://www.epi.mcgill.ca/hanley/c607/ch07/mm_ch_07.pdf

What if, in the sleep deprivation study, one used the (more natural ?) percent rather than absolute difference in reaction times?

And what if one were worried that the ‘formal’ and ‘official’ ‘Normality’ conditions for the validity of a t -test (emphasized by Baldi and Moore, only to be considerably relaxed in the end of the Chapter) were not satisfied? After all, it is difficult to imagine that in this population of truck-drivers, or more broadly, the distribution of absolute differences and the distribution of percent differences could **both** be **Normal** (Gaussian).

```
IndividualPctDiffFromDay0 = c(4, -8, -2, -7, -1, 3, 2, 4, 13, 0, -2, -5, 20, 35, -1, 4, 1, 2)
bootstrap <- do(1000) * mean( resample(IndividualPctDiffFromDay0) )
```

```
summary(bootstrap$mean)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
-2.611 1.696 3.134 3.370 4.779 12.236
confint(bootstrap) [extra -- silly-- decimal places removed, and %'s added]
 name lower upper level method estimate
1 mean -0.7% 8.5% 0.95 percentile 3.4%
[COMMENT: CI still crosses zero, just as do the CIs based on absolute differences]
```

authors report the SE of (or a CI for) each of the 2 means, but they are of no use: we aren’t interested in the means *per se*, but in the mean difference. Using $\text{Var}(\bar{y}_1 - \bar{y}_0) = \text{Var}(\bar{y}_1) + \text{Var}(\bar{y}_0)$ assumes *one* set of 18 subjects for the Day₀, and a *different* set of 18 for the Day₁ condition, a noisy contrast. See more on this in B&M, bottom of p. 424.

⁴The 2nd involved yields from barley grown from ‘regular’ and ‘kiln-dried’ seeds.

⁵See <http://www.epi.mcgill.ca/hanley/bios601/Mean-Quantile/First.t.test.pdf>

Power and sample size calculations e.g., : Is this milk watered down?

(Adapted from Q 15.17 from Moore and McCabe, 4th Edition)

A cheese maker buys milk from several suppliers. It suspects that some suppliers are adding water to their milk to increase their profits.

Excess water can be detected by measuring the freezing point of the ‘liquid’. The freezing temperature of natural milk varies according to a Gaussian distribution, with mean $\mu = -0.540^\circ$ Celsius (C) and standard deviation $\sigma = 0.008^\circ$ C. Added water raises the freezing temperature toward 0° C, the freezing point of water. The laboratory manager measures the freezing temperature of five consecutive lots of ‘milk’ from one supplier. The mean of these 5 measurements is -0.533° C. Is this good evidence that the producer is adding water to the milk?

Moore and McCabe asked students to ‘State hypotheses, carry out the test, give the P -value, and state your conclusion.’

In this course, we will go further and ask (Q1) how much water a farmer/supplier could add to the milk before (s)he has a 10% , 50%, 80% chance of getting caught (of the buyer ‘detecting’ the cheating). Assume the buyer continues to use an n of 5, and the same $\sigma = 0.008^\circ$ C, and bases the boundary for rejecting/accepting the product on a 5% ‘ α ’, and a 1-sided test, i.e, $z = 1.645$ ⁶, i.e., the buyer sets the cutoff⁷ at $-0.540 + (\text{qnorm}(0.95) = 1.645) \times 0.008/\sqrt{5} = -0.534^\circ$ C.

Assume that mixtures of M% milk and W% water would freeze at a mean of $\mu = (M/100) \times -0.545^\circ\text{C} + (W/100) \times 0^\circ\text{C}$ and that the σ would remain unchanged. Thus, mixtures of 99% milk and 1% water would freeze at a mean of $\mu = (99/100) \times -0.540^\circ\text{C} + (1/100) \times 0^\circ\text{C} = -0.5346^\circ\text{C}$. Mixtures of 98:2 would freeze at $\mu = -0.5292^\circ\text{C}$. [Hint: drawing overlapping distributions will help.]

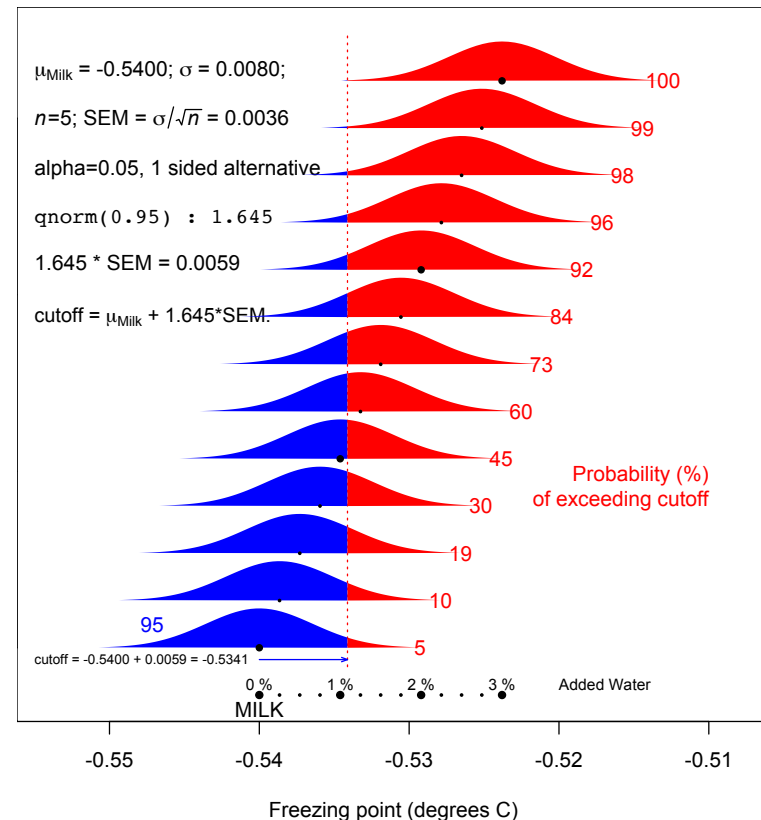
⁶Given the extensive experience with the variability of freezing point measurements, and the practice of using a ‘known’ σ in quality control, the use of a z multiple of 1.645 (rather than a t_4 multiple of 2.13) to establish the ‘critical value’ makes sense. It does not make sense to use an s based on a sample of 5 to re-estimate σ .

⁷http://ansci.illinois.edu/static/ansc438/Milkcompsynth/milkcomp_freezing.html states that ‘The current official freezing point limit (-0.525 degrees Horvet or -0.505 degrees C; see Sherbon 1988 for discussion of Horvet vs Centigrade) was designed for whole-herd, bulk-tank samples or processed milk samples, and not for samples from individual cows or individual quarters. The value of -0.525 degrees Horvet is considered the upper limit which statistically is suppose to be a cut-off for most, but not absolutely all, samples to be considered “water-free” (that is, no added water). <https://van.physics.illinois.edu/qa/listing.php?id=1606> states that ‘the exact freezing point of milk (also called the melting point) varies slightly according to the individual cow, the breed, the time of day / season that the milk is collected, the type of feed that the cow receives, etc. According to ... , the majority of cows produce milk with a natural freezing point of -0.5250 to -0.5650 C, with an average of about -0.5400 C.’

Once we have answered this further-reaching question, we will focus on a fixed ‘delta’, corresponding to a 99:1 mix of mild and water, and ask (Q2) what the chances are of detecting cheating if the buyer uses samples $n=10, 15$ or 20 rather than just 5 measurements.

Finally, again focusing on a fixed ‘delta’, corresponding to a 99:1 mix of mild:water, we ask (Q3): **at what ‘ n ’** does the chance of detecting cheating reach **80%**, a commonly used (but arbitrary) criterion used in sample-size planning by investigators seeking funding for their proposed research?

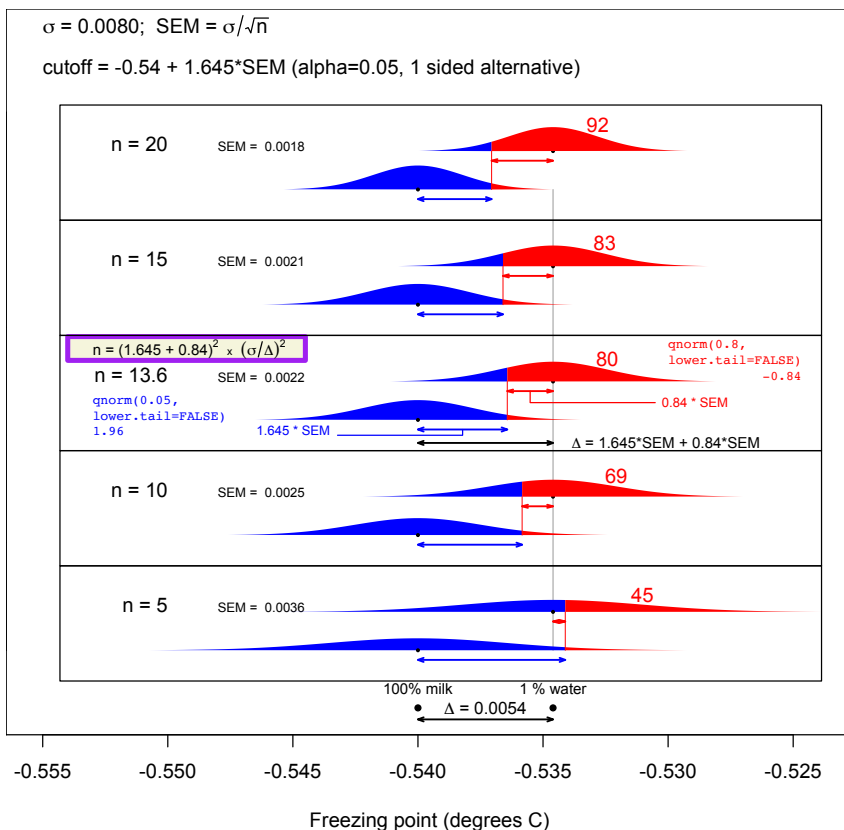
Q1: The calculations shown at the left below are used to set the cutoff; it is based on the null distribution shown at the bottom. Clearly the bigger the signal (the ‘ Δ ’) the more chance the test will ‘raise the red flag.’ It is 92% when it is a 98:2, and virtually 100% when it is a 97:3 mix.



The probabilities in red were calculated using the formula:
 $\text{pnorm}(\text{cutoff}, \text{mean} = \mu.\text{mixture}, \text{sd} = \text{SEM}, \text{lower.tail}=\text{FALSE})$

Q2: Suppose even a 1% added water is serious, and worth detecting. Clearly, from the previous Figure, and again at the bottom row of the Figure here, one has only a 45% chance of detecting it: there is a large overlap between the sampling distributions under the null (100% Milk) and the mixture (99% milk, 1% water) scenarios.

So, to better discriminate, one needs to make a bigger resting effort, and measure more lots, i.e., increase the n .



The larger n narrows and concentrates the sampling distribution. The width is governed by the SD of the sampling distribution of the mean of n measurements, i.e., by the Standard Error of the Mean, or $SEM = \sigma/\sqrt{n}$.

Because the null sampling distribution narrows, the cutoff is brought closer to the null. And under the alternative (non-null) scenario, a greater portion of its sampling distribution is to the right of (i.e., exceeds) the cutoff.

Indeed, under the alternative (i.e., cheating) scenario $n = 10$ the probability of exceeding the threshold is almost 70% when $n = 10$, 82% when $n = 15$ and 92% when $n=20$. You can check these for yourself in R using this expression:

```
pnorm(cutoff, mean = mu.mixture, sd = sigma/sqrt(n), lower.tail=FALSE)
```

Q3: If⁸ it makes sense to aim for a particular probability, then it is easy to see, from the middle panel, how to come up with a closed form formula that (a) allows you to compute the sample size ‘by hand’ and (b) shows you, more explicitly than the diagram or R code can, what drives the n .

The ‘balancing formula’, in SEM terms, is simply the n where

$$1.645 \times SEM + 0.84 \times SEM = \Delta.$$

Replacing each of the SEMs (assumed equal, because we assumed the variability is approx. the same under both scenarios) by σ/\sqrt{n} , i.e.,

$$1.645 \times \sigma/\sqrt{n} + 0.84 \times \sigma/\sqrt{n} = \Delta.$$

and solving for n , one gets

$$n = (1.645 + 0.84)^2 \times \left\{ \frac{\sigma}{\Delta} \right\}^2 = (1.645 + 0.84)^2 \times \left\{ \frac{Noise}{Signal} \right\}^2.$$

Notice the ‘anatomy’ or ‘structure’ of the formula. The *first* component has to do with the operating characteristics or performance of the test, i.e., the ‘type I error’ probability ‘ α ’⁹ and the desired power¹⁰ (the complement of the ‘type II error’ probability, β).

The *second* has to do with the context in which it is applied, i.e., the size of the ‘noise’ relative to the ‘signal.’ In our example, where the ‘Noise-to-Signal Ratio’ is $\frac{\sigma=0.0080}{\Delta=0.0054} = 1.48$, so that its square is 1.48^2 or approx 2.2, and $(1.645 + 0.84)^2 = 2.485^2 =$ approx 6.2,

$$n = 6.2 \times 2.2 = 13.6, \text{ approx, or, rounded up, } n = 14.$$

⁸What is magic or sacred about 80%?

⁹Had one set the the ‘type I error’ probability at ‘ α ’ = 0.01, the (one-sided) *z* value used in the cutoff would increase to $qnorm(0.99, lower.tail=TRUE) = 2.326$, instead of 1.645.. If it would make sense to use a *two sided* alternative, and an ‘ α ’ of 0.05, the Z would have been the familiar $qnorm(0.975, lower.tail=TRUE) = 1.96$ and -1.96 below.

¹⁰The *z* = 0.84 is for 80% power; for 50% power, use $qnorm(0.50, lower.tail=TRUE) = 0$; for 90% power, use *z* = $qnorm(0.90, lower.tail=TRUE) = 1.28$;

An **important point about interpreting p-values from statistical tests** – and the careful ‘legal’ wording of the probabilities shown in red.

In *Moore and McCabe*’s example, an $n = 5$ gives an SE of $\sigma/\sqrt{5} = 0.0080/2236 = 0.0036$ approx. So the cutoff for a 1 sided test with ‘ α ’ = 0.05 is 1.645×0.0036 or 0.0059 approx. above -0.5400, i.e., at -0.5341. This is computed under the null (innocence) hypothesis, namely that what we are testing is pure milk, with no added water. The 1 sided alternative is that we are testing a ‘less than 100%, more than 0%’ mix, where the mean is above (to right of) -0.540, i.e., on the (upper) ‘added water’ side of the null. Formally, these two hypotheses are

$$H_0 : \mu = -0.540; H_{alt} : \mu > -0.540.$$

Since the mean of the 5 measurements, namely -0.533°C, is to the right of (exceeds) this threshold, it would be considered ‘statistically significant at the 0.05 level.’ The actual p-value is `pnorm(-0.533, mean=-0.54, sd = 0.0036, lower.tail=FALSE) = 0.026`.

M&M had asked ask you to consider whether ‘**Is this good evidence that the producer is adding water to the milk?** : ‘State the hypotheses, carry out the test, give the P -value, and state your conclusion.’

And it is here that you **need to be nuanced; do not ‘jump to conclusions’ and immediately accuse the supplier of cheating.**

In particular, it would not be appropriate – or accurate – to say that you are $1 - 0.026 = 0.974 = 97.4\%$ certain that the supplier is cheating. Remember that a p-value is a probability concerning the data, conditional on (i.e., computed under the assumption that) H_0 being (is) true. In other words, the p-value has to do with $P(\text{data} \mid \text{‘innocence’})$, whereas at issue is the reverse, $P(\text{‘innocence’} \mid \text{data})$.¹¹

As to this latter probability (of being innocent), there are a lot of other factors to consider first, before accusing the supplier of cheating.

First, did you (re-)check the calculations? How recently was the instrument calibrated? etc.¹²

Second, why did you chose to test **this** supplier? Is it someone that the manager suspected based on previous data, or based on knowing that he

¹¹And, by the way, it won’t help to switch the conversation to a ‘Confidence Interval’ in the hope of simplifying the matter, and avoiding getting tangled up in statistical jargon, trying to explain all these concepts to the farmer.

¹²JH calls these ‘Type III errors’: the data were wrong, or the instrument was wrong, or the technician mis-calculated something. It should remind us that, in the *real* world, there are *many* alternative hypotheses, not just the one.

is behind in his loan payments to the bank? Or maybe the laboratory manager merely asked a technician to start randomly testing, and the first supplier (blindly) chosen was the manager’s brother-in-law?

So, you can see that, just as in medical tests, there are **many other pieces of evidence** or information, or circumstances, besides the p-value, that bear on the probability of innocence or guilt.¹³ This is very nicely brought out in the article ‘Are all p-values created equal?’ which you can here: <http://www.biostat.mcgill.ca/hanley/BionanoWorkshop/AreAllSigPValuesCreatedEqual.pdf>

Sadly, the mixing up of $P(\text{data} \mid \text{hypothesis})$ and $P(\text{data} \mid \text{data})$ – often referred to as ‘The Prosecutor’s Fallacy’ – is common, and can lead to serious harm. See the second-last page in this link <http://www.epi.mcgill.ca/hanley/IntMedResidents/P-Values.pdf>

The use of p-values works well in Quality Control, where the aim is to detect (the few) deviations (‘bad’ ones) from the desired specifications, to stop and fix the offending machine, or to flag defective batches. It is not clear that it is equally effective at identifying the (few) truly active (‘good’) compounds via the mass testing of lots of compounds, most of which are expected to be inactive – and then investing all one’s effort in these few ‘good’ ones at the next stage of development.

[Legalese – small print:] Possible **appropriate** wordings to accompany a $p=0.026$:

When (if) we test samples of pure milk, only 2.6% of test results are/would be this high or higher.

IF the only factor operating here were sampling variation, only 2.6% of test results on pure milk would be this high or higher.

Many shorten the latter to ‘It would be rare to get this high or a higher test value by chance (or by chance alone), (dangerously) using ‘*by*’ instead of ‘*if*.’ Adding the ‘*alone*’ does help direct attention to *what else* might be responsible for the extreme value.

Others go further astray: ‘The probability *that this result was due to chance* (or chance alone) is only 2.6%.’ This latter, quite misleading, statement should be avoided and ‘not used in better families.’ **Best** get in the habit of including the word ‘**if**’ in the statement, i.e., making it clear the statement is conditional. And avoid sliding over into the mistake – as the second statement does – of addressing the probability that the null (or any alternative) hypothesis is true / operating. That probability (about a hypothesis) is outside the scope of frequentist statistics, and should be left to experts in the substantive field. A non-expert would not pronounce on the probability that a patient has a certain condition, based *only* on the asterisks (the p-value) printed beside a single lab-chemistry value.

¹³Repeating the test is a common practice when the test result does not agree/fit with the other evidence.