017 - Statistical Power

EPIB 607

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca

slides compiled on October 15, 2021

Power and Sample Size

How much water a supplier could add to the milk before they have a 10% , 50%, 80% chance of getting caught, i.e., of the buyer detecting the cheating ?

Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples n=10, 15 or 20 rather than just 5 measurements?

At what n does the chance of detecting cheating reach 80%?

Another example on power: Lake Wobegon

Is this milk watered down?[1]

- 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.

[1] Adapted from Q 15.17 from Moore and McCabe, 4th Edition

Is this milk watered down?[1]

- 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 $\mu = -0.540°$ Celsius (C) and standard deviation $\sigma = 0.008°$C.
- Added water raises the freezing temperature toward $0°$C, the freezing point of water.

[1] Adapted from Q 15.17 from Moore and McCabe, 4th Edition

Is this milk watered down?[1]

- 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 $\mu = -0.540°$ Celsius (C) and standard deviation $\sigma = 0.008°$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.

[1] Adapted from Q 15.17 from Moore and McCabe, 4th Edition

Is this milk watered down?[1]

- 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 $\mu = -0.540°$ Celsius (C) and standard deviation $\sigma = 0.008°$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.
- Question: Is this good evidence that the producer is adding water to the milk?

[1] Adapted from Q 15.17 from Moore and McCabe, 4th Edition

Is this milk watered down?

- State hypotheses:

Is this milk watered down?

- State hypotheses:
  - $H_0 : \mu = -0.540°\text{C}$

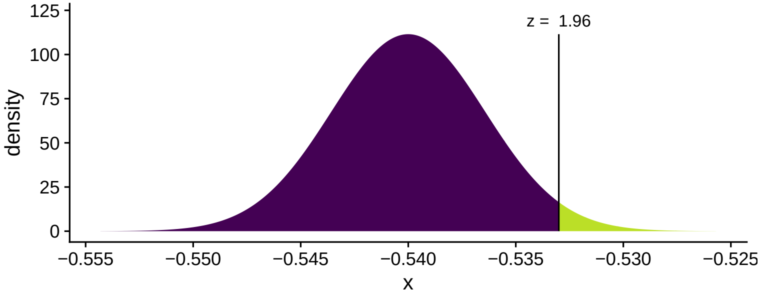
Is this milk watered down?

- State hypotheses:
  - $H_0 : \mu = -0.540°\mathrm{C}$
  - $H_a : \mu > -0.540°\mathrm{C}$
- Which test should we use and why?

Is this milk watered down?

- State hypotheses:
  - $H_0 : \mu = -0.540°\text{C}$
  - $H_a : \mu > -0.540°\text{C}$
- Which test should we use and why?

```
mosaic::xpnorm(q = -0.533, mean = -0.540, sd = 0.008/sqrt(5))
```




```
## [1] 0.9748004
```

Testing using the p-value

Appropriate wordings to accompany $p = 0.0252$:

- If we test samples of pure milk, only 2.6% of test results would be this high or higher.

Testing using the p-value

Appropriate wordings to accompany $p = 0.0252$:

- If we test samples of pure milk, only 2.6% of test results would be this high or higher.

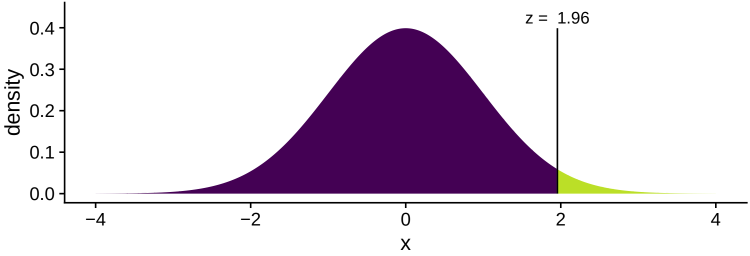
- IF the only factor operating here were sampling variation, only 2.6% of test results on pure milk would be this high or higher.

Test using a Z statistic

- $H_0 : \mu = -0.540°\text{C} \qquad H_a : \mu > -0.540°\text{C}$
- We can also standardize our observed mean and calculate the p-value under a $\mathcal{N}(0, 1)$

```
SEM <- 0.008/sqrt(5)
z_stat <- (-0.533 - (-0.540)) / SEM
mosaic::xpnorm(q = z_stat, mean = 0, sd = 1)

##
## If X ~ N(0, 1), then
## P(X <= 1.957) = P(Z <= 1.957) = 0.9748
## P(X >  1.957) = P(Z >  1.957) = 0.0252
##
```



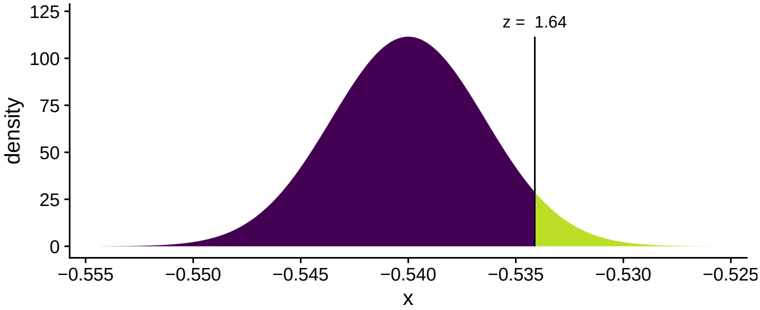

```
## [1] 0.9748004
```

Test using critical values

- An observed mean freezing temperature greater than -0.5341 rejects the null hypothesis:

```
mosaic::xqnorm(p = 0.95, mean = -0.540, sd = 0.008/sqrt(5))
```

```
##
## If X ~ N(-0.54, 0.003577709), then
## P(X <= -0.5341152) = 0.95
## P(X >  -0.5341152) = 0.05
##
```




```
## [1] -0.5341152
```

Test using critical values

```
mosaic::xqnorm(p = 0.95,
mean = -0.540,
sd = 0.008/sqrt(5))
```

```
mosaic::xpnorm(q = -0.533,
mean = -0.540,
sd = 0.008/sqrt(5))
```

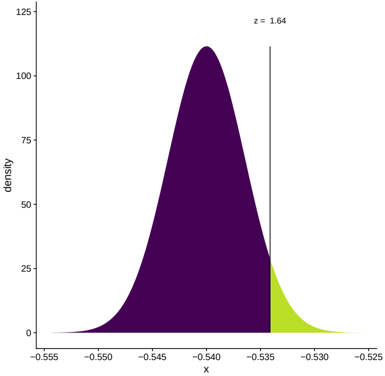


Figure: critical value under the null distribution

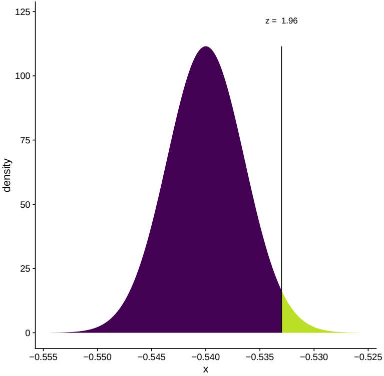


Figure: test statistic under the null distribution

Thus we reject H_0 at $\alpha = 0.05$.

Type I and II errors

What does it mean to reject H_0 at level α?

Type I and II errors

What does it mean to reject H_0 at level α?

- It means that, if H_0 were true and the procedure (sampling data, performing the significance test) were repeated many times, the testing procedure would reject H_0 $\alpha 100\%$ of the time.

		Truth about the population	
		H_0 true	H_a true
Decision based on sample	Reject H_0	Type I error	Correct decision
	Accept H_0	Correct decision	Type II error

Type I and II errors

In this special setting we give special names to the false positive and false negative rates:

- **Type I error (α):** probability that a significance test will reject H_0 when in fact H_0 is true.

Type I and II errors

In this special setting we give special names to the false positive and false negative rates:

- **Type I error (α):** probability that a significance test will reject H_0 when in fact H_0 is true.
- **Type II error (β):** probability that a significance test will fail to reject H_0 when H_0 is not true.

Type I and II errors

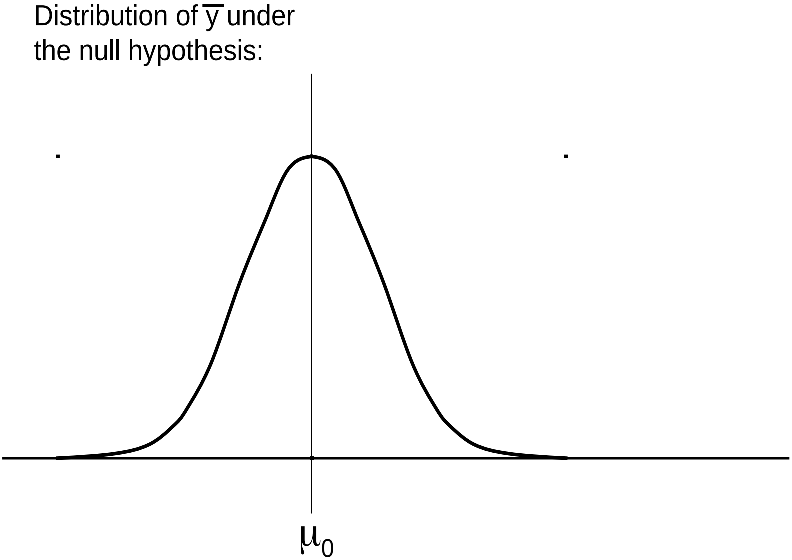
In this special setting we give special names to the false positive and false negative rates:

- **Type I error (α):** probability that a significance test will reject H_0 when in fact H_0 is true.
- **Type II error (β):** probability that a significance test will fail to reject H_0 when H_0 is not true.

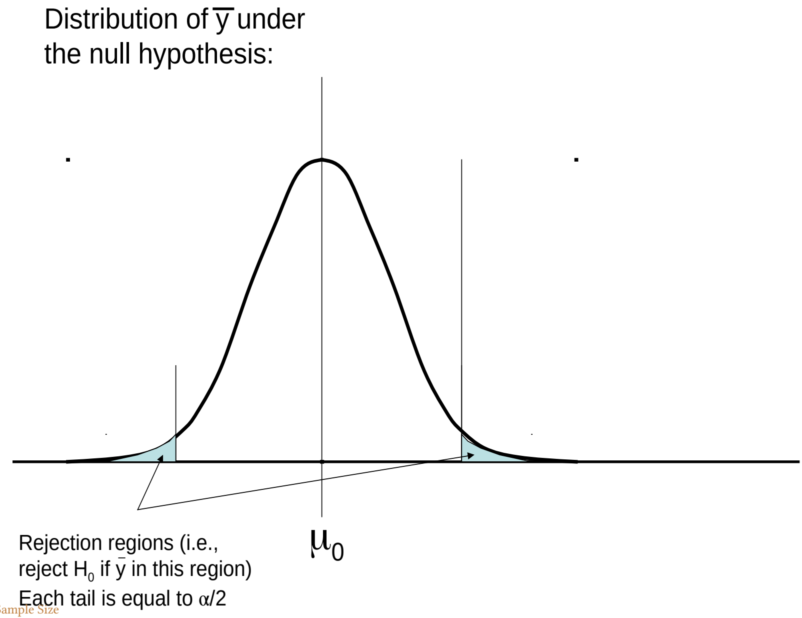
The Type I error is the significance level of the test, α, which is often set to 0.05.

As we will see in a moment, the Type II error, β, is determined by the sample size and the chosen Type I error rate/significance level. (Therefore, with α fixed at, say 0.05, the only way to reduce β is to increase n or decrease s.)

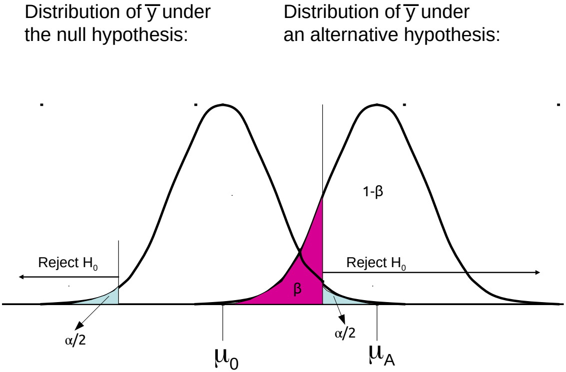
Type I and II errors



Type I and II errors

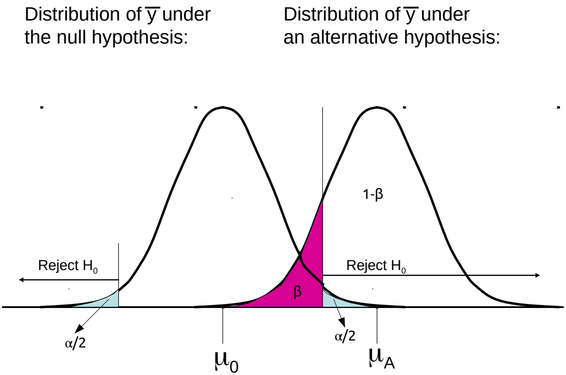


Type I and II errors



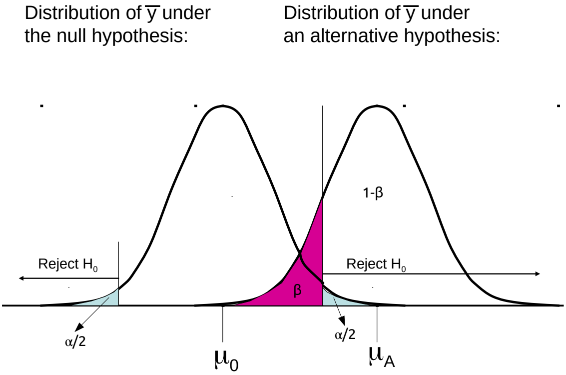
Type I and II errors

- The blue area represents the Type I error – the probability of rejecting H_0 **if** H_0 **is true**.



Type I and II errors

- The blue area represents the Type I error – the probability of rejecting H_0 **if H_0 is true**.
- The purple area represents the Type II error – the probability of *not* rejecting H_0 **if H_A is in fact true** (and therefore H_0 should be rejected).

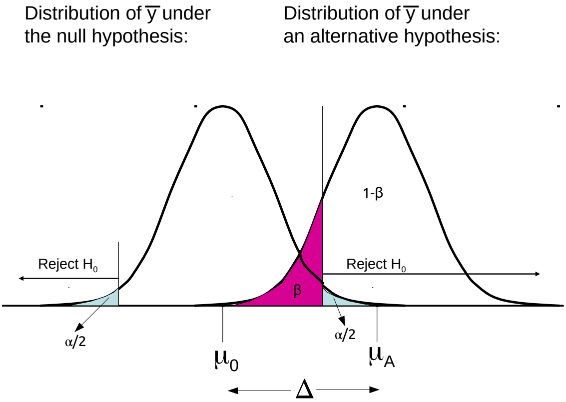


Type I and II errors

- Notice the distribution of the alternative has a different center, but the same SD

Type I and II errors

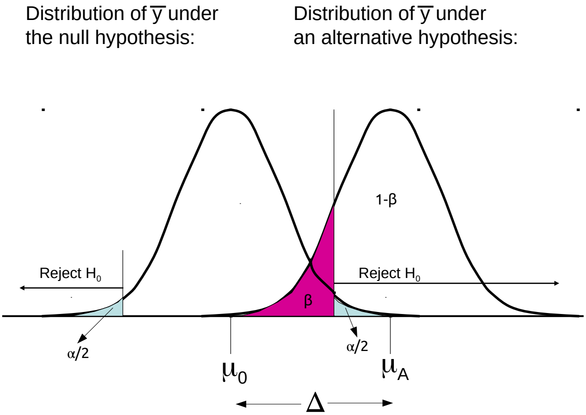
- Notice the distribution of the alternative has a different center, but the same SD
- The distance between μ_0 and the true value of μ (in our previous slide we called this μ_A) will affect the Type II error. This distance is denoted as Δ.



Power = $1 - \beta$

Definition 1 (Power = $1 - \beta$).

The probability that a fixed level α significance test will reject H_0 when a particular alternative value of the parameter is true is called the **power** of the test to detect the alternative.



Power and Sample Size: 3 questions

1. How much water a supplier could add to the milk before they have a 10% , 50%, 80% chance of getting caught, i.e., of the buyer detecting the cheating ?

Power and Sample Size: 3 questions

1. How much water a supplier could add to the milk before they have a 10% , 50%, 80% chance of getting caught, i.e., of the buyer detecting the cheating ?

2. Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples n=10, 15 or 20 rather than just 5 measurements?

Power and Sample Size: 3 questions

1. How much water a supplier could add to the milk before they have a 10% , 50%, 80% chance of getting caught, i.e., of the buyer detecting the cheating ?

2. Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples n=10, 15 or 20 rather than just 5 measurements?

3. 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*)

Power and Sample Size

How much water a supplier could add to the milk before they have a 10% , 50%, 80% chance of getting caught, i.e., of the buyer detecting the cheating ?

Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples n=10, 15 or 20 rather than just 5 measurements?

At what n does the chance of detecting cheating reach 80%?

Another example on power: Lake Wobegon

Statistical Power: the chance of getting caught

- We want to know how much water a farmer could add to the milk before they have a 10% , 50%, 80% chance of getting caught (of the buyer detecting the cheating).
- Assume the buyer continues to use an $n = 5$, and the same $\sigma = 0.008°\text{C}$, and bases the boundary for rejecting/accepting the product on a $\alpha = 0.05$, and a 1-sided test which translates to the buyer setting the cutoff at

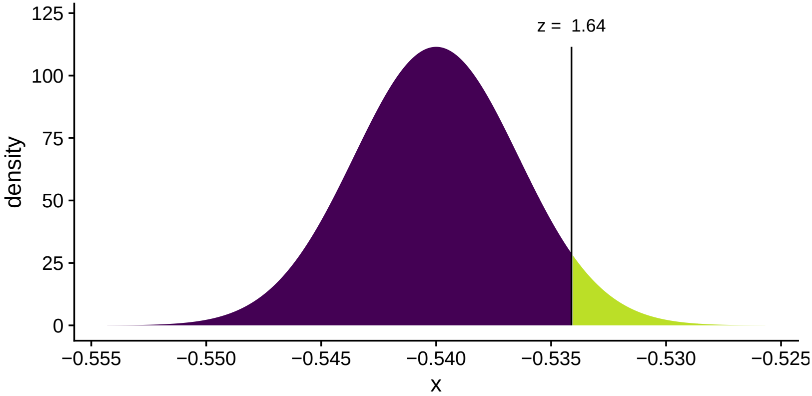
$$-0.540 + 1.645 \times 0.008/\sqrt{5} = -0.534°\text{C}.$$

- This is equivalent to `qnorm(p = 0.95, mean = -0.540, sd = 0.008/sqrt(5))`

The cutoff at $\alpha = 0.05$

- $-0.540 + 1.645 \times 0.008/\sqrt{5} = -0.534°\text{C}.$

```
mosaic::xqnorm(p = 0.95, mean = -0.540, sd = 0.008/sqrt(5))
```




```
## [1] -0.5341152
```

Statistical Power

- Assume that mixtures of M% milk and W% water would freeze at a mean of

$$\mu_{mixture} = (M/100) \times -0.545^{\circ}C + (W/100) \times 0^{\circ}C$$

and that the σ would remain unchanged.

Statistical Power

- Assume that mixtures of M% milk and W% water would freeze at a mean of

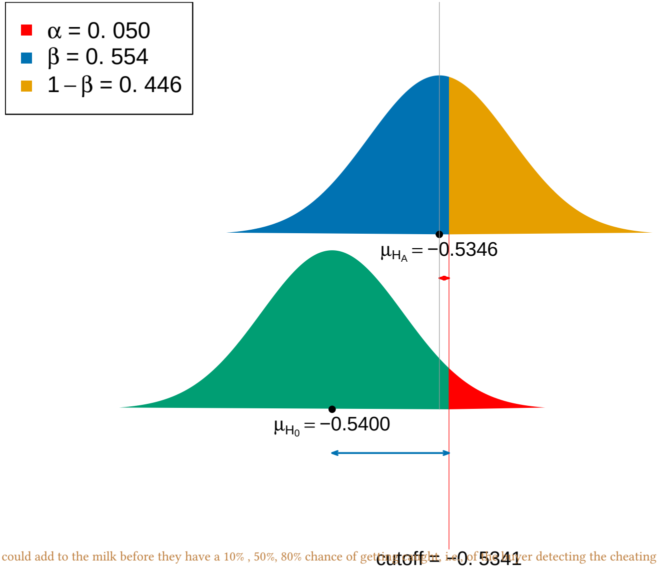
$$\mu_{mixture} = (M/100) \times -0.545^{\circ}C + (W/100) \times 0^{\circ}C$$

  and that the σ would remain unchanged.

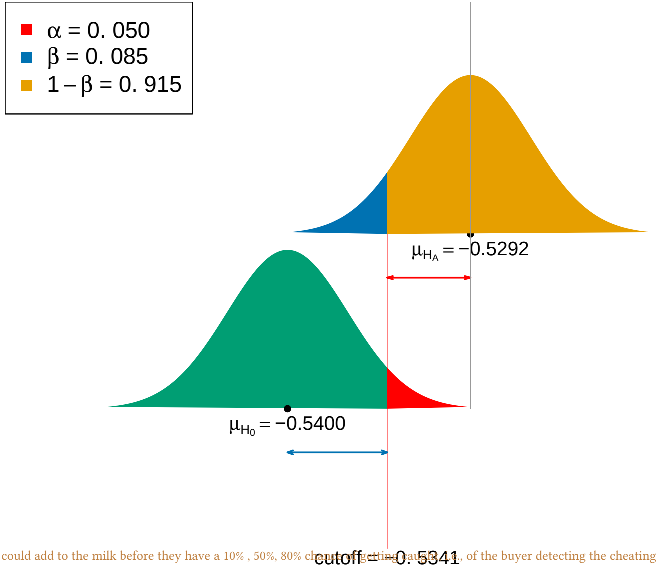
- Thus, mixtures of 99% milk and 1% water would freeze at a mean of $\mu = (99/100) \times -0.540^{\circ}C + (1/100) \times 0^{\circ}C = -0.5346^{\circ}C$.

% milk	% water	mean (μ)
99	1	-0.5346$^{\circ}C$
98	2	-0.5292$^{\circ}C$
97	3	-0.5238$^{\circ}C$

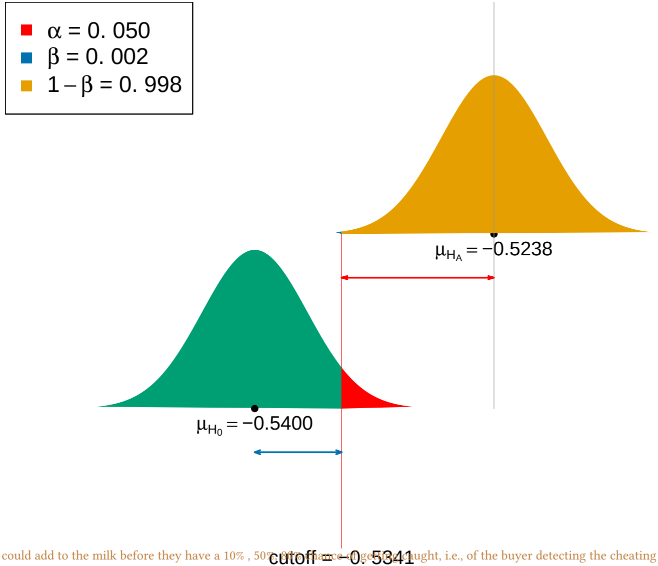
If the supplier added 1% water to the milk

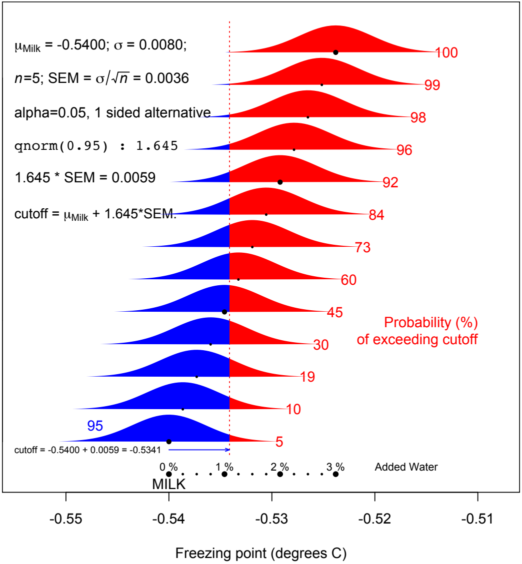


If the supplier added 2% water to the milk



If the supplier added 3% water to the milk





The probabilities in red were calculated using the formula: `stats::pnorm(cutoff, mean = mu.mixture, sd = SEM, lower.tail=FALSE)`

Statistical Power: the chance of getting caught

- The calculations shown at the left in the figure on the previous slide are used to set the cutoff; it is based on the **null** distribution shown at the bottom.

Statistical Power: the chance of getting caught

- The calculations shown at the left in the figure on the previous slide 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.

Power and Sample Size

How much water a supplier could add to the milk before they have a 10% , 50%, 80% chance of getting caught, i.e., of the buyer detecting the cheating ?

Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples n=10, 15 or 20 rather than just 5 measurements?

At what n does the chance of detecting cheating reach 80%?

Another example on power: Lake Wobegon

Power as a function of sample size

- Suppose even a 1% added water is serious, and worth detecting.
- Clearly, from the previous Figure, and again at the bottom row of the following Figure, 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**.

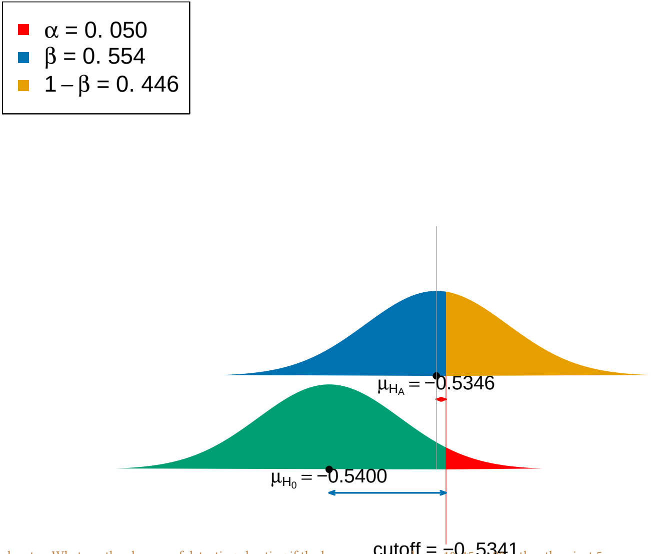
Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples n=10, 15 or 20 rather than just 5 measurements?

Power as a function of sample size

- Suppose even a 1% added water is serious, and worth detecting.
- Clearly, from the previous Figure, and again at the bottom row of the following Figure, 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.

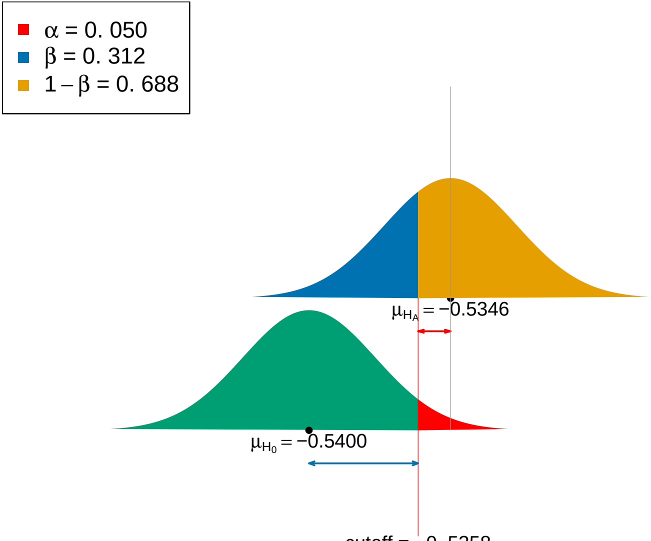
Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples n=10, 15 or 20 rather than just 5 measurements?

When the buyer uses samples of size 5



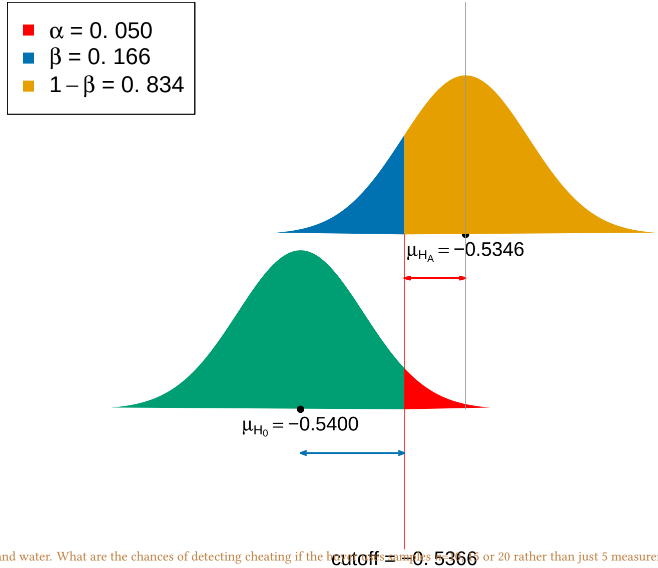
Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples n=10, 15 or 20 rather than just 5 measurements?

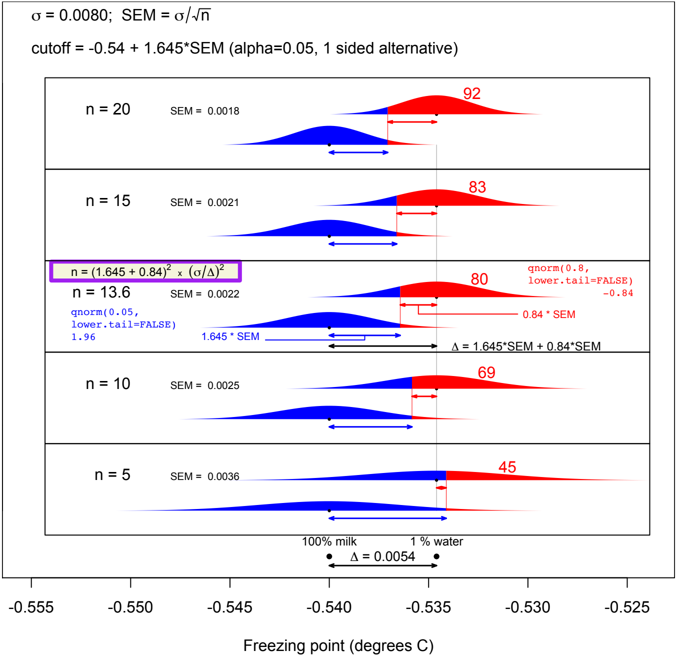
When the buyer uses samples of size 10



Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples n=10, 15 or 20 rather than just 5 measurements?

When the buyer uses samples of size 15





Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples n=10, 15 or 20 rather than just 5 measurements?

Increasing n leads to increased power

- 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 = $\sigma/\sqrt{n}$.

Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples n=10, 15 or 20 rather than just 5 measurements?

Increasing n leads to increased power

- 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.

Increasing n leads to increased power

- 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 the probability of exceeding the threshold is almost 70% when $n = 10$, 82% when $n = 15$ and 92% when n=20.

Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples n=10, 15 or 20 rather than just 5 measurements?

Increasing n leads to increased power

- 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 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:

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

Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples n=10, 15 or 20 rather than just 5 measurements?

Power and Sample Size

How much water a supplier could add to the milk before they have a 10% , 50%, 80% chance of getting caught, i.e., of the buyer detecting the cheating ?

Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples n=10, 15 or 20 rather than just 5 measurements?

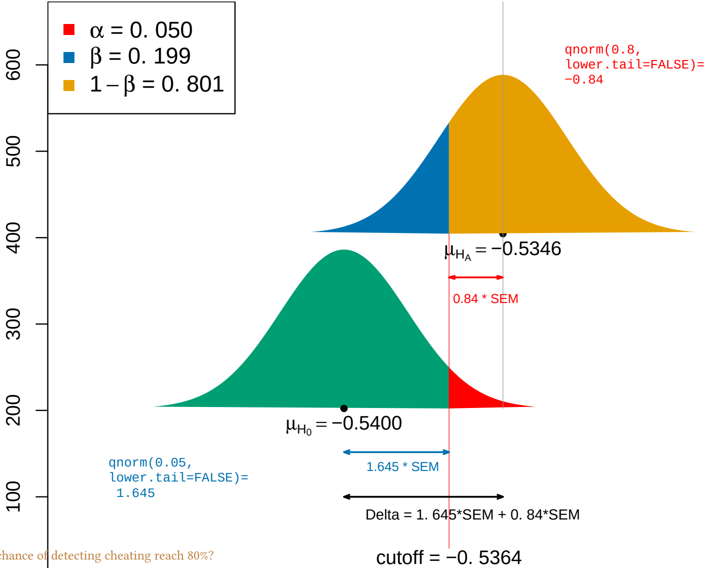
At what n does the chance of detecting cheating reach 80%?

Another example on power: Lake Wobegon

What sample size needed?

- We can 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



What sample size needed?

- 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 $\sigma/\sqrt{n}$, i.e.,

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

and solving for n, one gets

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

What sample size needed?

- Notice the 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 and the desired power (the complement of the type II error probability, β).

What sample size needed?

- Notice the 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 and the desired power (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.

What sample size needed?

- Notice the 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 and the desired power (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,

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

Code for null and alternative distribution plots

```
source("https://raw.githubusercontent.com/sahirbhatnagar/EPIB607/master/inst/code/plot_null_alt.R")

mu0 <- -0.540 # mean under the null
mha <- 0.99*-0.540 # mean under the alternative
s <- 0.0080 # sample/population SD
n <- 5 # sample size
cutoff <- mu0 + qnorm(0.95) * s / sqrt(n)

power_plot(n = n,
s = s,
mu0 = mu0,
mha = mha,
cutoff = cutoff,
alternative = "greater",
xlab = "Freezing point (degrees C)")
```

Interpreting p-values from statistical tests

- In the milk example, an $n = 5$ gives an SEM of $\sigma/\sqrt{5}$ = 0.0080/2236 = 0.0036. So the cutoff for a 1 sided test with α = 0.05 is $1.645 \times 0.0036 = 0.0059$ 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.

Interpreting p-values from statistical tests

- In the milk example, an $n = 5$ gives an SEM of $\sigma/\sqrt{5}$ = 0.0080/2236 = 0.0036. So the cutoff for a 1 sided test with α = 0.05 is $1.645 \times 0.0036 = 0.0059$ 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

$$\mathrm{H}_0 : \mu = -0.540; \; H_{alt} : \mu > -0.540.$$

Interpreting p-values from statistical tests

- In the milk example, an $n = 5$ gives an SEM of $\sigma/\sqrt{5}$ = 0.0080/2236 = 0.0036. So the cutoff for a 1 sided test with α = 0.05 is $1.645 \times 0.0036 = 0.0059$ 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

$$\mathrm{H}_0 : \mu = -0.540; \; H_{alt} : \mu > -0.540.$$

- 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)` = <u>0.026</u>.

Is this good evidence that the producer is adding water to the milk?

Is this good evidence that the producer is adding water to the milk?

- Do not jump to conclusions and immediately accuse the supplier of cheating

Is this good evidence that the producer is adding water to the milk?

- 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(data | 'innocence'), whereas at issue is the reverse, P('innocence' | data).

Is this good evidence that the producer is adding water to the milk?

- 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(data | 'innocence'), whereas at issue is the reverse, P('innocence' | data).
- As to this latter probability (of being innocent), there are a lot of other factors to consider first, before accusing the supplier of cheating.

Other factors to consider before accusing the supplier

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

Other factors to consider before accusing the supplier

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

Other factors to consider before accusing the supplier

- First, did you (re-)check the calculations? How recently was the instrument calibrated? etc.
- JH 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.
- Second, why did you chose to test **this** supplier?

Other factors to consider before accusing the supplier

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

Are all p-values created equal?

- 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.
- 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(data | hypothesis) and P(data | data) – often referred to as 'The Prosecutor's Fallacy' – is common, and can lead to serious harm.

When does the p-value work well?

- 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.

When does the p-value work well?

- 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.

Power and Sample Size

How much water a supplier could add to the milk before they have a 10% , 50%, 80% chance of getting caught, i.e., of the buyer detecting the cheating ?

Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples n=10, 15 or 20 rather than just 5 measurements?

At what n does the chance of detecting cheating reach 80%?

Another example on power: Lake Wobegon

Power: Lake Wobegon

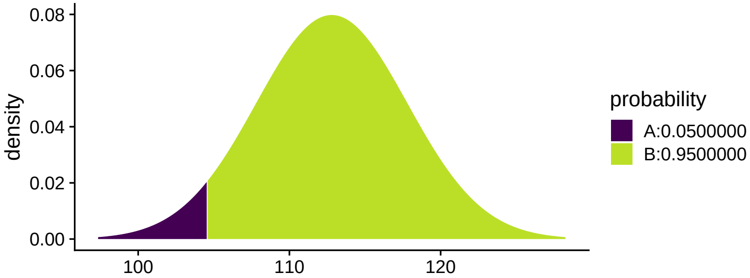
- It is claimed that the children of Lake Wobegon are above average. Take a simple random sample of 9 children from Lake Wobegon, and measure their IQ to obtain a sample mean of 112.8.
- IQ scores are scaled to be Normally distributed with mean 100 and standard deviation 15.

1. Null and alternative hypotheses: The claim made is that the population of children Lake Wobegon have higher than average intellidence. Thus the null hypothesis is that the population has average intelligence, or a score of 100. Therefore $H_0 : \mu = 100$, and the (one-sided) alternative is $H_A : \mu > 100$.

Lower limit of 95% CI

1. Hypotheses. $H_0 : \mu = 100$, $H_A : \mu > 100$.
2. Calculate 95% CI.

```
mosaic::xqnorm(p = c(0.05,1), 112.8, 15/sqrt(9))
```

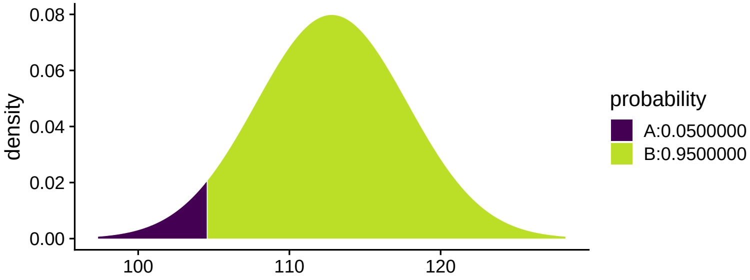



```
## [1] 104.5757      Inf
```

Lower limit of 95% CI

1. Hypotheses. $H_0 : \mu = 100$, $H_A : \mu > 100$.
2. Calculate 95% CI.

```
mosaic::xqnorm(p = c(0.05,1), 112.8, 15/sqrt(9))
```




```
## [1] 104.5757      Inf
```

3. Statement. The lower limit of the CI excludes $\mu_0 = 100$, and so there is evidence to suggest that the children at Lake Wobegon are brighter than other children at the $\alpha = 0.05$ level.

Power

Steps to finding power:

1. State null hypothesis, H_0, and state a specific alternative, H_A, as the minimum (clinical/substantive) departure from the null hypothesis that would be of interest.

Power

Steps to finding power:

1. State null hypothesis, H_0, and state a specific alternative, H_A, as the minimum (clinical/substantive) departure from the null hypothesis that would be of interest.
2. Find the values of $\overline{y}$ that lead to rejection of H_0.

Power

Steps to finding power:

1. State null hypothesis, H_0, and state a specific alternative, H_A, as the minimum (clinical/substantive) departure from the null hypothesis that would be of interest.
2. Find the values of $\overline{y}$ that lead to rejection of H_0.
3. Calculate the probability of observing the values found in (2) when the alternative is true.

Power

Example: Lake Wobegon
Suppose you hope to use a **one-sided** test to show that the children from Lake Wobegon are at least 10 points higher than average on the IQ test. What power do you have to detect this with the sample of 9 children if using a 0.05-level test?

Power

Example: Lake Wobegon
Suppose you hope to use a **one-sided** test to show that the children from Lake Wobegon are at least 10 points higher than average on the IQ test. What power do you have to detect this with the sample of 9 children if using a 0.05-level test?

1. Hypotheses. $H_0 : \mu = 100$, $H_A : \mu > 110$.

Power

Example: Lake Wobegon
Suppose you hope to use a **one-sided** test to show that the children from Lake Wobegon are at least 10 points higher than average on the IQ test. What power do you have to detect this with the sample of 9 children if using a 0.05-level test?

1. Hypotheses. $H_0 : \mu = 100$, $H_A : \mu > 110$.

2. Find values of the sample mean that reject the null.

   The test will reject H_0 at the 0.05 level whenever

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{y} - 100}{15/\sqrt{9}} \geq 1.645.$$

   Now we must translate this back to values of $\bar{y}$...

Power

Example: Lake Wobegon

2. Values of the sample mean that reject H_0 (con't)

   The test will reject H_0 at the 0.05 level whenever

   $$\frac{\bar{y} - 100}{15/\sqrt{9}} \geq 1.645,$$

   which means we reject H_0 whenever

   $$\bar{y} \geq 1.645 \times 15/\sqrt{9} + 100 = 108.2$$

   If H_0 is true, the probability of seeing an IQ score as big as 108.2 or bigger is 5%.

Power

Example: Lake Wobegon

3. Find the probability of rejecting H_0 if $\mu = \mu_A = 110$.

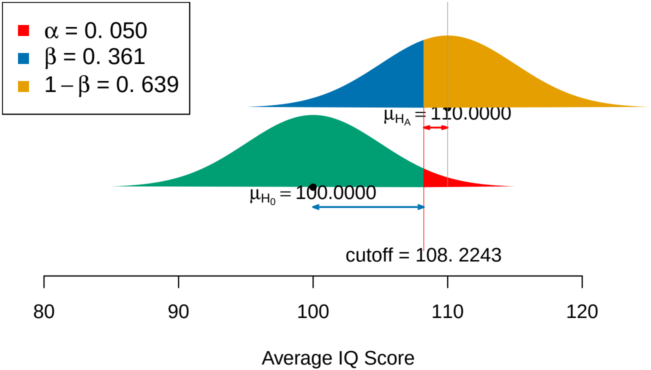
$P(\bar{y} > 108.2 | \mu = \mu_A = 110)$

$$
\begin{aligned}
&= P\left(\frac{\bar{y} - \mu_A}{\sigma/\sqrt{n}} > \frac{108.2 - 110}{15/\sqrt{9}} \middle| \mu = \mu_A = 110 \right) \\
&= P(z > -0.36) \\
&= 0.64
\end{aligned}
$$

So there is approximately a 2/3 chance of detecting a difference of 10 points on the IQ scale at the 0.05 level of significance with a sample size of 9.

Null and alternative distribution plots

```
source("https://raw.githubusercontent.com/sahirbhatnagar/EPIB607/master/inst/code/plot_null_alt.R")
power_plot(n = 9, s = 15, mu0 = 100, mha = 110,
cutoff = 100 + qnorm(0.95) * 15 / sqrt(9),
alternative = "greater", xlab = "Average IQ Score")
```



Power

Example: Lake Wobegon

If you hoped to use a two-sided test to show that the children from Lake Wobegon are at least 10 points higher than average on the IQ test, what power do you have with the sample size of 9 and a 0.05-level test?

Power

Example: Lake Wobegon

If you hoped to use a two-sided test to show that the children from Lake Wobegon are at least 10 points higher than average on the IQ test, what power do you have with the sample size of 9 and a 0.05-level test?

1. Hypotheses. $H_0 : \mu = 100$, $H_A : \mu = 110$.

Power

Example: Lake Wobegon

If you hoped to use a two-sided test to show that the children from Lake Wobegon are at least 10 points higher than average on the IQ test, what power do you have with the sample size of 9 and a 0.05-level test?

1. Hypotheses. $H_0 : \mu = 100$, $H_A : \mu = 110$.

2. Find values of the sample mean that reject the null.

   The test will reject H_0 at the 0.05 level whenever

$$z = \frac{\overline{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{\overline{y} - 100}{15/\sqrt{9}} \geq 1.96 \qquad \textbf{OR when}$$

$$z = \frac{\overline{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{\overline{y} - 100}{15/\sqrt{9}} \leq -1.96$$

   since we are performing a two-sided test.

Power

Thus we reject H_0 if $\overline{y} \geq 1.96 \times 5 + 100 = 109.8$ **OR** if $\overline{y} \leq -1.96 \times 5 + 100 = 90.2$

3. Find the probability of rejecting H_0 if $\mu = \mu_A = 110$.

$$P(\overline{y} > 109.8 \textbf{ OR } \overline{y} < 90.2 | \mu = \mu_A = 110)$$

$$\begin{aligned}
&= P(\overline{y} > 109.8 | \mu_A = 110) + P(\overline{y} < 90.2 | \mu_A = 110) \\
&= P\left(\frac{\overline{y} - \mu_A}{\sigma/\sqrt{n}} > \frac{109.8 - 110}{15/\sqrt{9}} \middle| \mu_A = 110 \right) \\
&\quad + P\left(\frac{\overline{y} - \mu_A}{\sigma/\sqrt{n}} < \frac{90.2 - 110}{15/\sqrt{9}} \middle| \mu_A = 110 \right) \\
&= P(z > -0.04) + P(z < -3.96) \\
&= 0.52 + 3.7 \times 10^{-5} \approx 0.52
\end{aligned}$$

There is about a 1/2 chance of detecting a difference of 10 pts on the IQ scale at the 0.05 level of significance with $n = 9$ using a two-sided alternative hypothesis.

Power

- Steps 1 and 2 used to find the power of a one-sided test to detect a difference of x points above (or below) the population mean are similar to the steps for finding the power of a two-sided test to detect a difference of x points on either side of the population mean.
- However for a two-sided test, there will be two sets of values of $\overline{y}$ that lead us to reject H_0 (also, z_α for one-sided and $z_{\alpha/2}$ for two-sided).
- **However**, there is one critical difference in the third step: we need to calculate the probability of seeing $\overline{y}$ in either of the two tails (rejection regions) of the null distribution under the assumption that the true distribution has mean μ_A. So there will be two probabilities to calculate in the third step.

Power

- Steps 1 and 2 used to find the power of a one-sided test to detect a difference of x points above (or below) the population mean are similar to the steps for finding the power of a two-sided test to detect a difference of x points on either side of the population mean.
- However for a two-sided test, there will be two sets of values of $\overline{y}$ that lead us to reject H_0 (also, z_α for one-sided and $z_{\alpha/2}$ for two-sided).
- **However**, there is one critical difference in the third step: we need to calculate the probability of seeing $\overline{y}$ in either of the two tails (rejection regions) of the null distribution under the assumption that the true distribution has mean μ_A. So there will be two probabilities to calculate in the third step.
- Note: If we felt that the minimum significant departure from μ_0 was different above and below (e.g., we are interested in increases of blood pressure of at least 3.5mmHg and decreases of blood pressure of at least 2mmHg), we perform the calculations as though we were interested in the minimum of the two values (Why?).

Power

Exercises: Lake Wobegon

Find the power of the following tests, assuming two-sided alternative hypotheses:

1. A 0.05-level test to detect a difference of 15 points on the IQ scale using the 9 children.
2. A 0.05-level test to detect a difference of 5 points on the IQ scale using the 9 children.
3. A 0.05-level test to detect a difference of 10 points on the IQ scale using 25 children.
4. A 0.01-level test to detect a difference of 10 points on the IQ scale using the 9 children.

Session Info

```
R version 4.1.1 (2021-08-10)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 21.04

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.13.so

attached base packages:
[1] tools     stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] DT_0.16           mosaic_1.7.0      Matrix_1.3-2      mosaicData_0.20.1
 [5] ggformula_0.9.4   ggstance_0.3.4    lattice_0.20-41   kableExtra_1.2.1
 [9] socviz_1.2        gapminder_0.3.0   here_0.1          NCStats_0.4.7
[13] FSA_0.8.30        forcats_0.5.1     stringr_1.4.0     dplyr_1.0.7
[17] purrr_0.3.4       readr_1.4.0       tidyr_1.1.3       tibble_3.1.5
[21] ggplot2_3.3.5     tidyverse_1.3.0   knitr_1.36

loaded via a namespace (and not attached):
 [1] fs_1.5.0           lubridate_1.7.9    webshot_0.5.2      httr_1.4.2
 [5] rprojroot_2.0.2    latex2exp_0.4.0    backports_1.2.1    utf8_1.2.2
 [9] R6_2.5.1           DBI_1.1.1          colorspace_2.0-2   withr_2.4.2
[13] tidyselect_1.1.1   gridExtra_2.3      leaflet_2.0.3      curl_4.3.2
[17] compiler_4.1.1     cli_3.0.1          rvest_1.0.0        pacman_0.5.1
[21] xml2_1.3.2         ggdendro_0.1.22    labeling_0.4.2     mosaicCore_0.8.0
[25] scales_1.1.1       digest_0.6.28      foreign_0.8-81     rmarkdown_2.11.3
[29] rio_0.5.16         pkgconfig_2.0.3    htmltools_0.5.2    highr_0.9
[33] dbplyr_1.4.4       fastmap_1.1.0      htmlwidgets_1.5.3  rlang_0.4.11
[37] readxl_1.3.1       rstudioapi_0.13    farver_2.1.0       generics_0.1.0
[41] jsonlite_1.7.2     crosstalk_1.1.1    zip_2.2.0          car_3.0-9
[45] magrittr_2.0.1     Rcpp_1.0.7         munsell_0.5.0      fansi_0.5.0
[49] abind_1.4-5        lifecycle_1.0.1    stringi_1.7.5      carData_3.0-4
[53] MASS_7.3-53.1      plyr_1.8.6         grid_4.1.1         blob_1.2.1
[57] ggrepel_0.8.2      crayon_1.4.1       cowplot_1.1.0      haven_2.3.1
[61] splines_4.1.1      hms_1.1.1          pillar_1.6.3       reprex_0.3.0
[65] glue_1.4.2         evaluate_0.14      data.table_1.14.2  modelr_0.1.8
[69] vctrs_0.3.8        tweenr_1.0.1       cellranger_1.1.0   gtable_0.3.0
[73] polyclip_1.10-0    assertthat_0.2.1   TeachingDemos_2.12 xfun_0.26
[77] ggforce_0.3.2      openxlsx_4.1.5     broom_0.7.2        viridisLite_0.4.0
```