p-values, Power and Sample Size

JH Notes: Inference about a Population Mean (μ)

Sahir Bhatnagar and James Hanley

EPIB 607
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca
https://sahirbhatnagar.com/EPIB607/

October 15, 2018

p-values

p-values and statistical tests

Definition 1 (p-value)

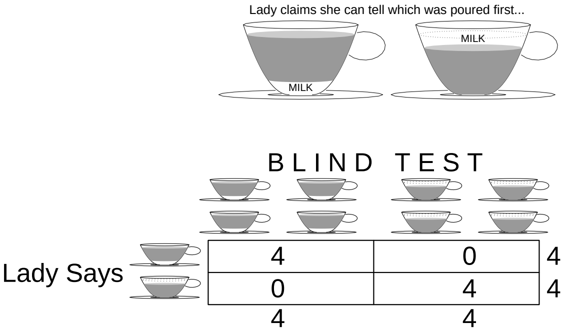
*A **probability concerning the observed data**, calculated under a **Null Hypothesis** assumption, i.e., assuming that the only factor operating is sampling or measurement variation.*

Use: To assess the evidence provided by the sample data in relation to a pre-specified claim or ‘hypothesis’ concerning some parameter(s) or data-generating process.

Basis: As with a confidence interval, it makes use of the concept of a *distribution*.

Caution: A p-value is NOT the probability that the null ‘hypothesis’ is true

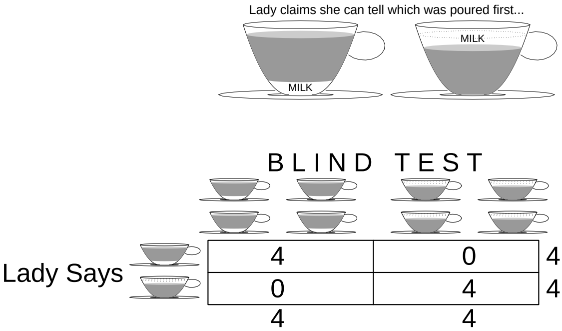
Example 1 – from *Design of Experiments*, by R.A. Fisher



Null Hypothesis (H_{null}): she can not tell them apart, i.e., just guessing.

Alternative Hypothesis (H_{alt}): she can.

Example 1 – from *Design of Experiments*, by R.A. Fisher

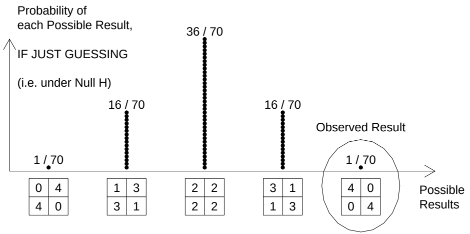


Null Hypothesis (H_{null}): she can not tell them apart, i.e., just guessing.

Alternative Hypothesis (H_{alt}): she can.

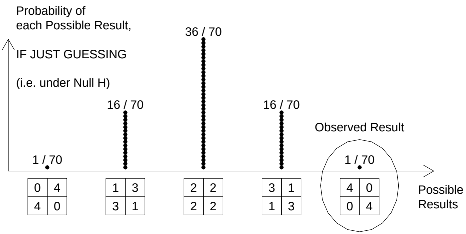
The evidence provided by the test

- Rank possible test results by degree of evidence against H_{null}.
- "*p*-value" is the probability, calculated under null hypothesis, of observing a result as extreme as, or more extreme than, the one that was obtained/observed.



The evidence provided by the test

- Rank possible test results by degree of evidence against H_{null}.
- "p-value" is the probability, calculated under null hypothesis, of observing a result as extreme as, or more extreme than, the one that was obtained/observed.

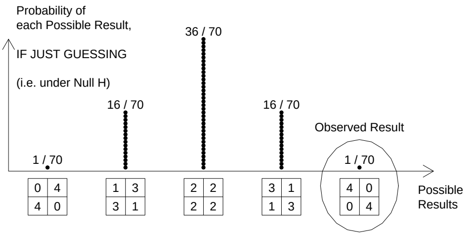


In this example, observed result is the most extreme, so

$P_{value} = \text{Prob[correctly identifying all 4, IF merely guessing]} = 1/70 = 0.014.$

The evidence provided by the test

- Rank possible test results by degree of evidence against H_{null}.
- "*p*-value" is the probability, calculated under null hypothesis, of observing a result as extreme as, or more extreme than, the one that was obtained/observed.



In this example, observed result is the most extreme, so

$$P_{value} = \text{Prob[correctly identifying all 4, IF merely guessing]} = 1/70 = 0.014.$$

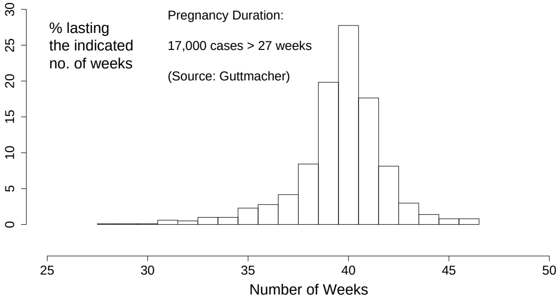
- Interpretation of such data often rather simplistic, as if these *data alone* should *decide*: i.e. if $P_{value} < 0.05$, we ~~'reject' H~~$_{null}$; if $P_{value} > 0.05$, we don't (or worse, we ~~'accept' H~~$_{null}$). Avoid such simplistic 'conclusions'.

Example 2 – Preston-Jones vs. Preston-Jones, English House of Lords, 1949

Divorce case:

- Sole evidence of adultery was that a baby was born almost 50 weeks after husband had gone abroad on military service. The appeal failed.
- To quote the court:
  - *"The appeal judges agreed that the limit of credibility had to be drawn somewhere, but on medical evidence 349 (days) while improbable, was scientifically possible."*

Example 2 – data collected from the 1970s



- *p*-value, calculated under "Null" assumption that husband was father, = 'tail area' or probability corresponding to an observation of '50 or more weeks' in above distribution
- Same system used to report how extreme a lab value is – are told where value is located in distribution of values from healthy (reference) population.

p-value via the Normal (Gaussian) distribution.

- When judging extremeness of a sample mean or proportion (or difference between 2 sample means or proportions) calculated from an amount of information that is sufficient for the Central Limit Theorem to apply, one can use Gaussian distribution to readily obtain the p-value.
- Calculate how many standard errors of the statistic, $SE_{statistic}$, the statistic is from where null hypothesis states true value should be. This "number of SE's" is in this situation referred to as a 'Z_{value}.'

$$Z_{value} = \frac{statistic - \text{its expected value under } H_{null}}{SE_{statistic}}.$$

  p-value can then be obtained by determining what % of values in a Normal distribution are as extreme or more extreme than this Z_{value}.
- If n is small enough that value of $SE_{statistic}$, is itself subject to some uncertainty, one would instead refer the "number of SE's" to a more appropriate reference distribution, such as Student's t- distribution.

More about the p-value

- The p-value is a **probability concerning data, conditional on the Null Hypothesis being true**.

More about the p-value

- The p-value is a **probability concerning data, conditional on the Null Hypothesis being true**.
- **Naive (and not so naive) end-users sometimes interpret the p-value as the probability that Null Hypothesis is true**, *conditional on – i.e. given – the data.*

More about the p-value

- The p-value is a **probability concerning data**, **conditional on the Null Hypothesis being true**.
- **Naive (and not so naive) end-users sometimes interpret the p-value as the probability that Null Hypothesis is true**, *conditional on – i.e. given – the data.*

$$\begin{aligned} p_{value} &= P(\text{this or more extreme data}|H_0) \\ &\neq P(H_0|\text{this or more extreme data}). \end{aligned}$$

- Statistical tests are often coded as statistically significant or not according to whether results are extreme or not with respect to a reference (null) distribution. But a test result is just one piece of data, and needs to be considered *along with rest of evidence* before coming to a 'conclusion.'
- **Likewise with statistical 'tests': the p-value is just one more piece of *evidence*, hardly enough to 'conclude' anything**.

The prosecutor’s fallacy [1]

- A criminal leaves fifty thousand blood cells at the scene of a crime which is just barely enough to stain a handkerchief.
- A forensic scientist extracts DNA from the sample to create a ‘DNA fingerprint’.
- Its pattern resembles that of a suspect.
- The scientist calculates that the chance of a match between the sample and a random member of the public is one in a million. How incriminating is this evidence?

[1] Who’s the DNA fingerprinting pointing at? New Scientist, 1994.01.29, 51-52.

The prosecutor’s fallacy

- Statistician Peter Donnelly opened a new area of debate, remarking that
  - Forensic evidence answers the question: “What is the probability that the defendant’s DNA profile matches that of the crime sample, assuming that the defendant is innocent?”
  - While the jury must try to answer the question: “What is the probability that the defendant is innocent, assuming that the DNA profiles of the defendant and the crime sample match?”

The prosecutor's fallacy

- Statistician Peter Donnelly opened a new area of debate, remarking that
  - Forensic evidence answers the question: "What is the probability that the defendant's DNA profile matches that of the crime sample, assuming that the defendant is innocent?"
  - While the jury must try to answer the question: "What is the probability that the defendant is innocent, assuming that the DNA profiles of the defendant and the crime sample match?"
- The error in mixing up these two probabilities is called "**the prosecutor's fallacy**," and it is suggested that newspapers regularly make this error.
- Donnelly's testimony convinced the judges that the case before them involved an example of this and they ordered a retrial.

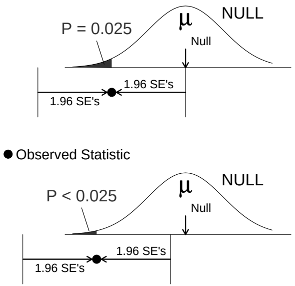
The prosecutor's fallacy in a game of poker

- Imagine the judges were playing a game of poker with the Archbishop of Canterbury.
- If the Archbishop were to deal a royal flush on the first hand, one might suspect him of cheating.
- The probability of the Archbishop dealing a royal flush on any one hand, assuming he is an honest card player, is about 1 in 70 000.
- But if the judges were asked whether the Archbishop was honest, given that he had just dealt a royal flush, they would be likely to quote a probability greater than 1 in 70 000.

The prosecutor's fallacy in a game of poker

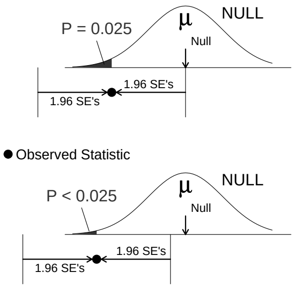
- Imagine the judges were playing a game of poker with the Archbishop of Canterbury.
- If the Archbishop were to deal a royal flush on the first hand, one might suspect him of cheating.
- The probability of the Archbishop dealing a royal flush on any one hand, assuming he is an honest card player, is about 1 in 70 000.
- But if the judges were asked whether the Archbishop was honest, given that he had just dealt a royal flush, they would be likely to quote a probability greater than 1 in 70 000.
- The first probability is analogous to the answer of the forensic scientist's question, and the second probability analogous to the answer of the jury's question.

(Intimate) Relationship between p-value and CI



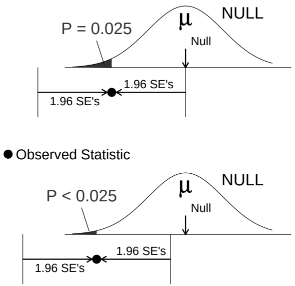
- (Upper graph) If upper limit of 95% CI *just touches* null value, then the 2 sided p-value is 0.05 (or 1 sided p-value is 0.025).

(Intimate) Relationship between *p*-value and CI



- (Upper graph) If upper limit of 95% CI *just touches* null value, then the 2 sided *p*-value is 0.05 (or 1 sided *p*-value is 0.025).
- (Lower graph) If upper limit *excludes* null value, then the 2 sided *p*-value is less than 0.05 (or 1 sided *p*-value is less than 0.025).

(Intimate) Relationship between *p*-value and CI



- (Upper graph) If upper limit of 95% CI *just touches* null value, then the 2 sided *p*-value is 0.05 (or 1 sided *p*-value is 0.025).
- (Lower graph) If upper limit *excludes* null value, then the 2 sided *p*-value is less than 0.05 (or 1 sided *p*-value is less than 0.025).
- (Graph not shown) If CI *includes* null value, then the 2-sided *p*-value is greater than (the conventional) 0.05, and thus observed statistic is “not statistically significantly different” from hypothesized null value.

Don’t be overly-impressed by p-values

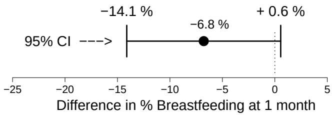
- p-values and ‘significance tests’ widely misunderstood and misused.
- Very large or very small n’s can influence what is or is not ‘statistically significant.’
- Use CI’s instead.
- *Pre study* power calculations (the chance that results will be ‘statistically significant’, as a function of the true underlying difference) of some help.
- *post-study* (i.e., *after the data have ‘spoken’*), a CI is much more relevant, as it focuses on magnitude & precision, not on a probability calculated under H_{null}.

Applications

Do infant formula samples ↓ $dur^{n.}$ of breastfeeding?[2]

Randomized Clinical Trial (RCT) which withheld free formula samples [given by baby-food companies to breast-feeding mothers leaving Montreal General Hospital with their newborn infants] from a random half of those studied.

At 1 month	Mothers given sample	not given sample	Total	**Conclusion...**
Still Breast feeding	175 (*77%*)	182 (*84%*)	357 (80.4%)	P=0.07. So, ... the difference is "Not Statistically Significant" at 0.05 level
Not Breast feeding	52	35	87	
Total	227	217	444	



[2] Bergevin Y, Dougherty C, Kramer MS. Lancet. 1983 1(8334):1148-51

Messages

- no matter whether the *p*-value is “statistically significant” or not, always look at the location and width of the confidence interval. it gives you a better and more complete indication of the magnitude of the effect and of the precision with which it was measured.
- this is an example of an **inconclusive negative** study, since it has **insufficient precision** (“resolving power”) **to distinguish** between two important possibilities – **no harm**, and what authoroties would consider a **substantial harm: a reduction of 10 percentage points** in breastfeeding rates .
- “**statistically** significant“ and “**clinically**-” (or “**public health**-”) significant are different concepts.
- (message from 1st author:) plan to have **enough statistical power**. his study had only 50% power to detect a difference of 10 percentage points)

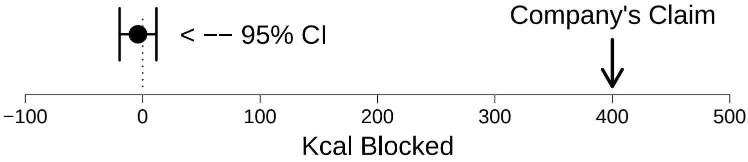
Do starch blockers really block calorie absorption?

Starch blockers – their effect on calorie absorption from a high-starch meal. Bo-Linn GW. et al New Eng J Med. 307(23):1413-6, 1982 Dec 2

- Known for more than 25 years that certain plant foods, e.g., kidney beans & wheat, contain a substance that inhibits activity of salivary and pancreatic amylase.
- More recently, this antiamylase has been purified and marketed for use in weight control under generic name "starch blockers."
- Although this approach to weight control is highly popular, it has never been shown whether starch-blocker tablets actually reduce absorption of calories from starch.
- Using a one-day calorie-balance technique and a high starch (100 g) meal (spaghetti, tomato sauce, and bread), we measured excretion of fecal calories after $n = 5$ normal subjects in a cross-over trial had taken either placebo or starch-blocker tablets.
- If the starch-blocker tablets had prevented the digestion of starch, fecal calorie excretion should have increased by 400 kcal.

Do starch blockers really block calorie absorption?

- However, fecal calorie excretion was same on the 2 test days (mean $\pm$ S.E.M., 80 $\pm$ 4 as compared with 78 $\pm$ 2).



- We conclude that starch blocker tablets do not inhibit the digestion and absorption of starch calories in human beings.
- EFFECT IS MINISCULE (AND ESTIMATE QUITE PRECISE) AND VERY FAR FROM COMPANY'S CLAIM !!!
- A '**DEFINITIVELY NEGATIVE**' STUDY.

Summary

SUMMARY - 1

- Confidence intervals preferable to p-values, since they are expressed in terms of (comparative) parameter of interest; they allow us to judge magnitude and its precision, and help us in 'ruling in / out' certain parameter values.

- A 'statistically significant' difference does not necessarily imply a clinically important difference.

- A 'not-statistically-significant' difference does not necessarily imply that we have ruled out a clinically important difference.

SUMMARY - 2

- Precise estimates distinguish b/w that which – if it were true – would be important and that which – if it were true – would not. ‘n’ an important determinant of precision.
- A lab value in upper 1% of reference distribution (of values derived from people without known diseases/conditions) does not mean that there is a 1% chance that person in whom it was measured is healthy; i.e., it doesn’t mean that there’s a 99% chance that the person in whom it was measured does have some disease/condition.
- Likewise, p-value $\neq$ probability that null hypothesis is true.
- The fact that

  $$Prob[the\ data \mid Healthy] \text{ is small [or large]}$$

  does not necessarily mean that

  $$Prob[Healthy \mid the\ data] \text{ is small [or large]}$$

SUMMARY - 3

- Ultimately, *p*-values, CI's and other evidence from a study need to be combined with other information bearing on parameter or process.
- Don't treat any one study as last word on the topic.

Power and Sample Size

Is this milk watered down?[3]

- 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.

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

Is this milk watered down?[3]

- 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.

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

Is this milk watered down?[3]

- 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.

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

Is this milk watered down?[3]

- 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?

[3] 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?

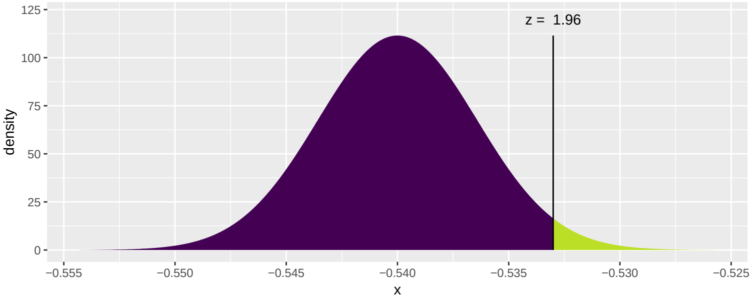
- State hypotheses:
  - $H_0 : \mu = -0.540$°C
  - $H_a : \mu > -0.540$°C
- Which test should we use and why?

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))
##
## If X ~ N(-0.54, 0.003578), then
## P(X <= -0.533) = P(Z <= 1.957) = 0.9748
## P(X >  -0.533) = P(Z >  1.957) = 0.0252
##
```

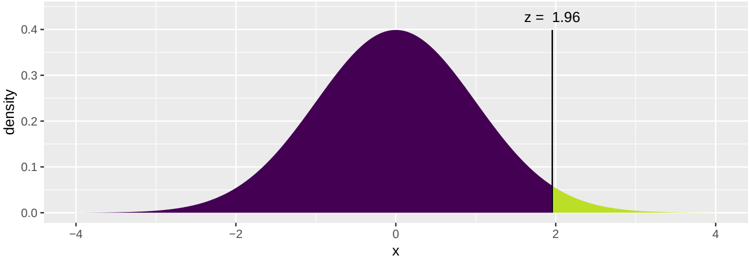


Test using a Z statistic

- $H_0 : \mu = -0.540°\text{C} \qquad H_a : \mu > -0.540°\text{C}$
- 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
##
```

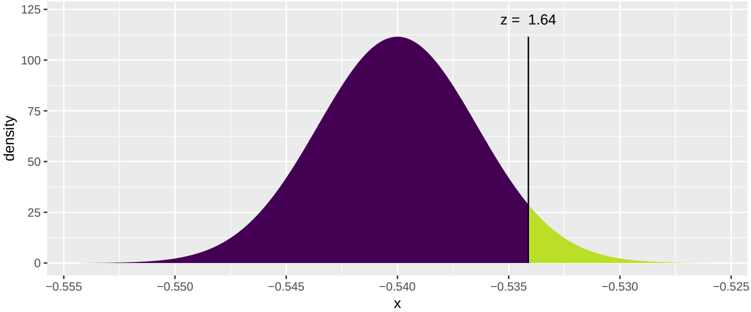


Test using critical values

- 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

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

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

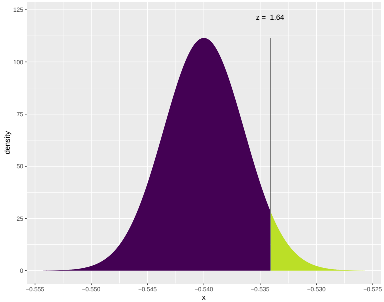


Fig.: critical value under the null distribution

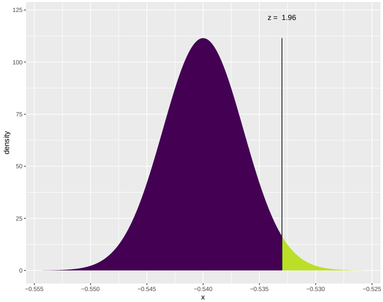


Fig.: test statistic under the null distribution

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.

Decision based on sample		Truth about the population: H_0 true	Truth about the population: H_a true
	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:

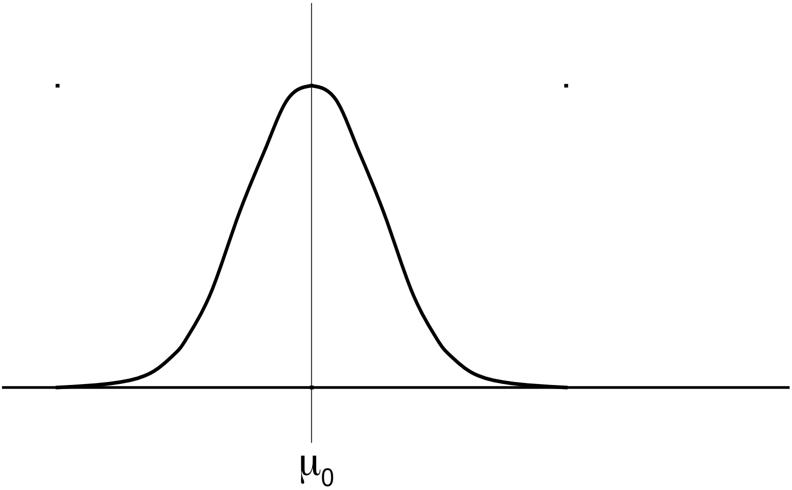
- **Type I error (α):** probability that a significance test will reject H_0 when in fact H_0 is true.
- **Type II error (β):** probability that a significance test will fail to reject H_0 when H_0 is not true.

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

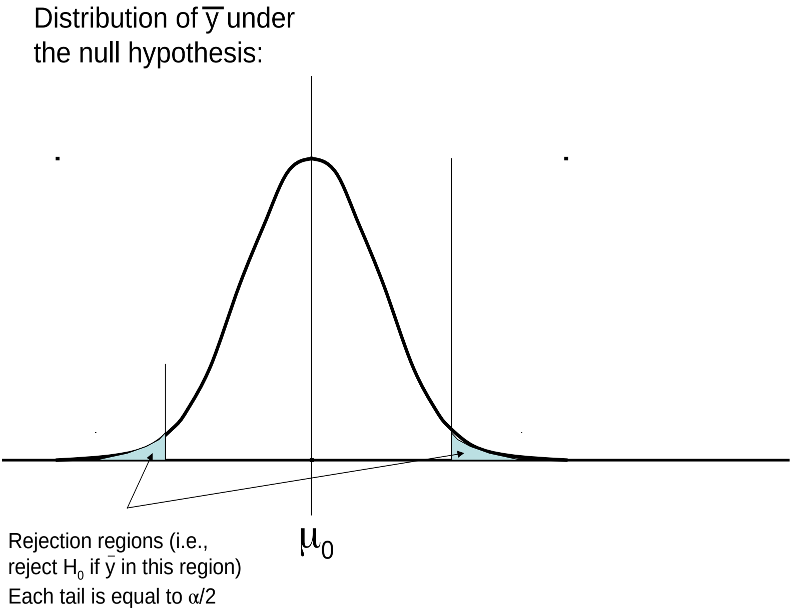
As we will see in a moment, the Type II error, β, is determined by the sample size and the chosen Type I error rate/significance level. (Therefore, with α fixed at, say 0.05, the only way to reduce β is to increase n or decrease s.)

Type I and II errors

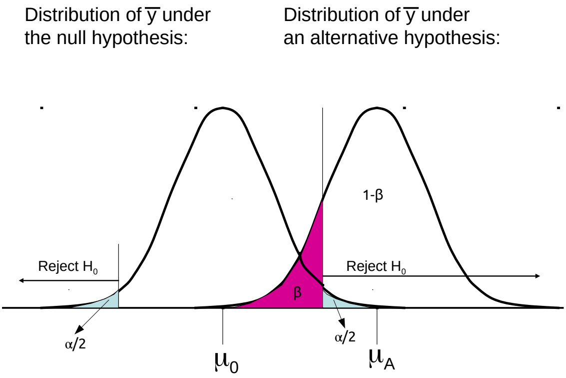
Distribution of $\overline{y}$ under the null hypothesis:



Type I and II errors

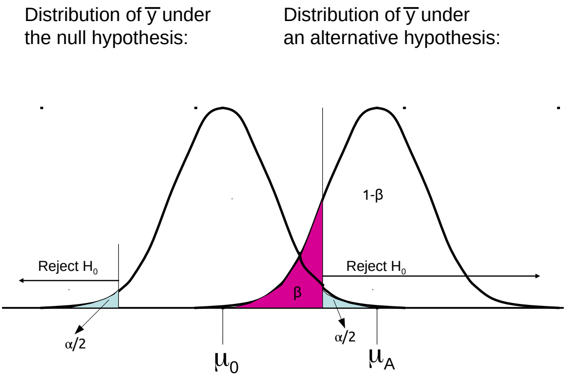


Type I and II errors



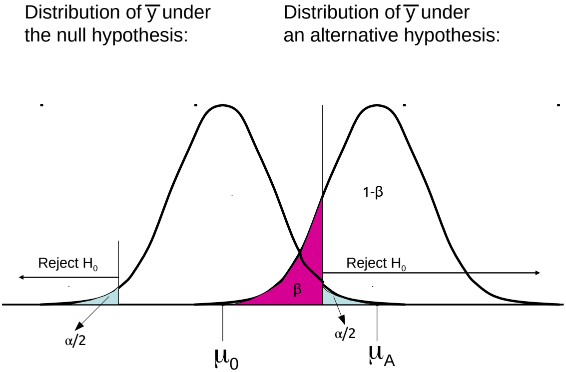
Type I and II errors

- 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).

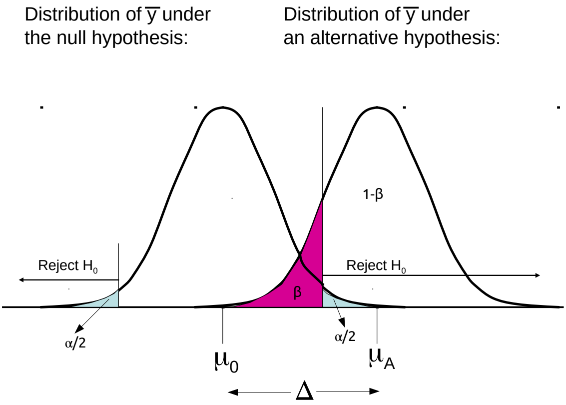


Type I and II errors

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

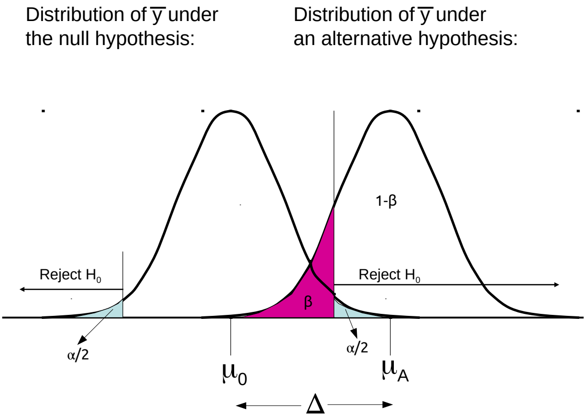
Type I and II errors

- 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 2 (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*)

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 ?

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 $\sigma = 0.008$°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

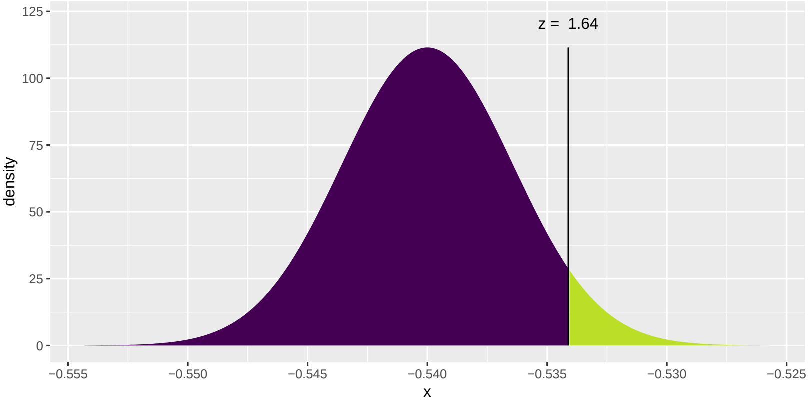
$$-0.540 + 1.645 \times 0.008/\sqrt{5} = -0.534°\text{C}.$$

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

The cutoff at $\alpha = 0.05$

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

```r
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

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

and that the σ would remain unchanged.

Statistical Power

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

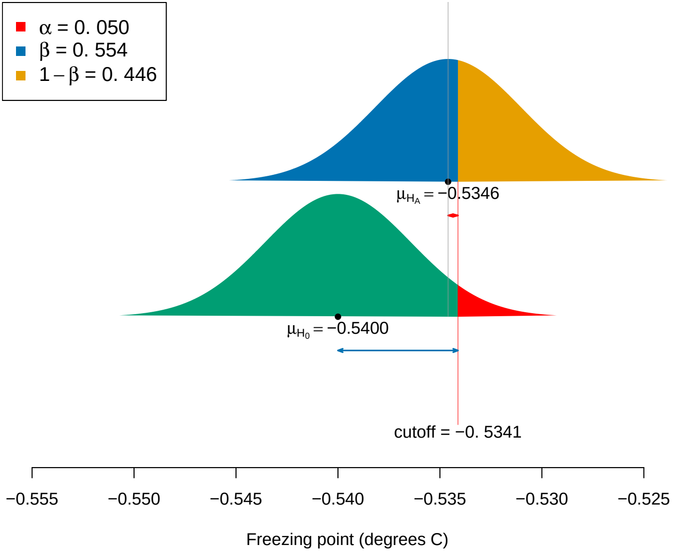
$$\mu_{mixture} = (M/100) \times -0.545°C + (W/100) \times 0°C$$

  and that the σ would remain unchanged.

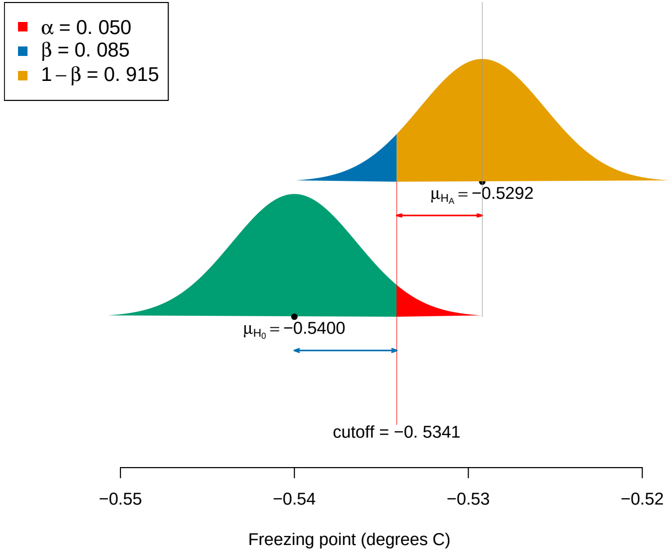
- Thus, mixtures of 99% milk and 1% water would freeze at a mean of $\mu = (99/100) \times -0.540°C + (1/100) \times 0°C = -0.5346°C$.

% milk	% water	mean (μ)
99	1	-0.5346°C
98	2	-0.5292°C
97	3	-0.5238°C

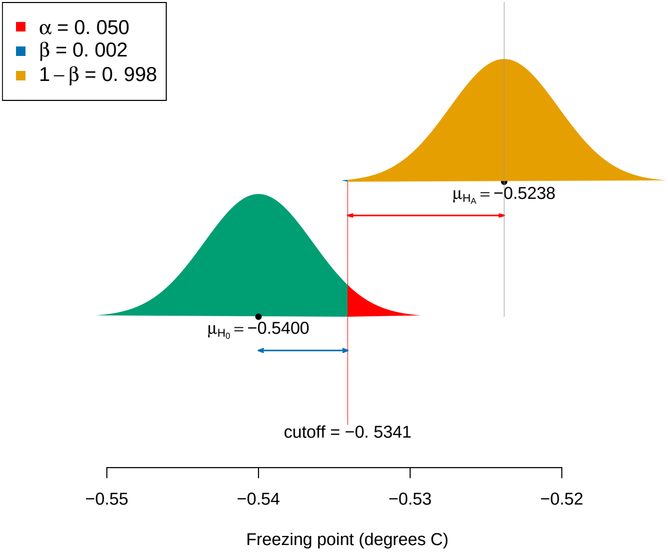
If the supplier added 1% water to the milk

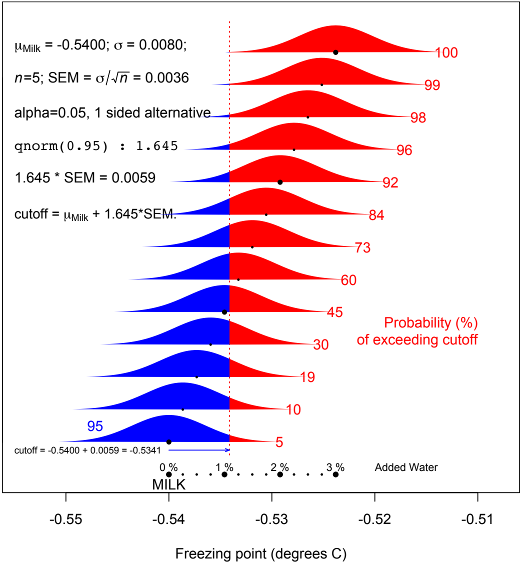


If the supplier added 2% water to the milk



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 <u>null</u> 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 <u>null</u> 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.

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?

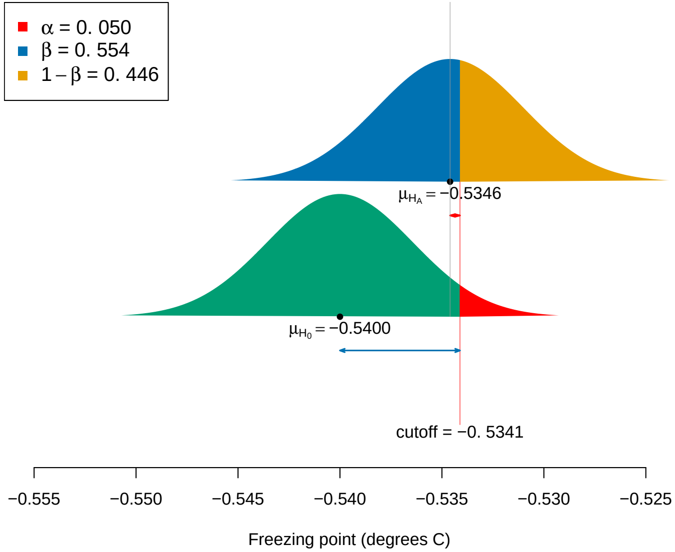
Power as a function of sample size

- 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**.

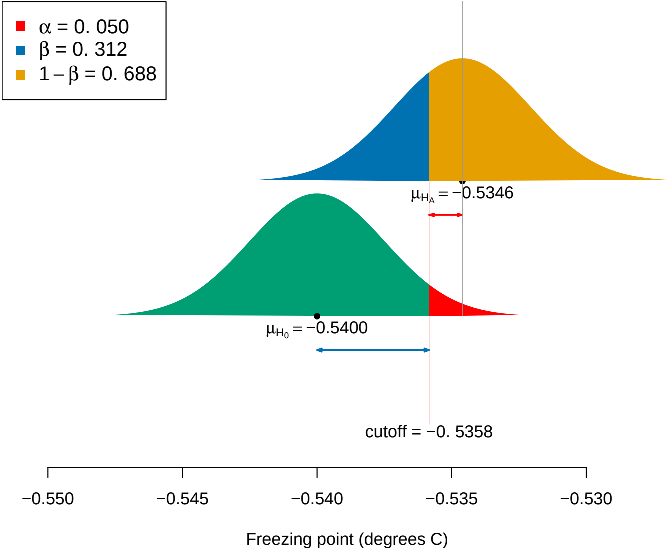
Power as a function of sample size

- 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*.

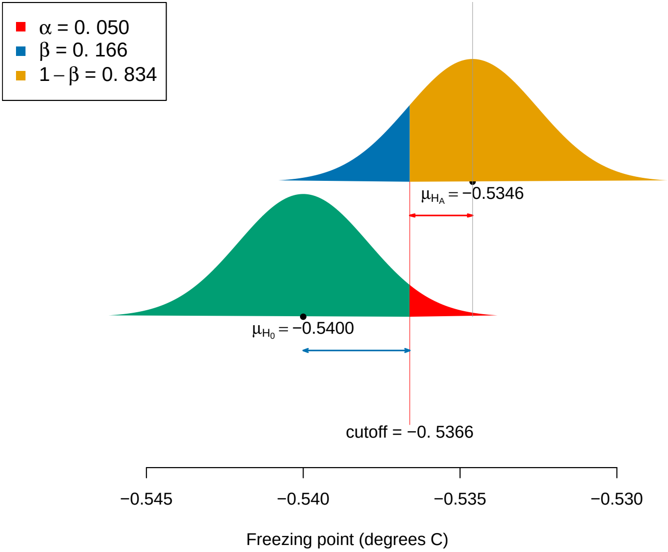
When the buyer uses samples of size 5

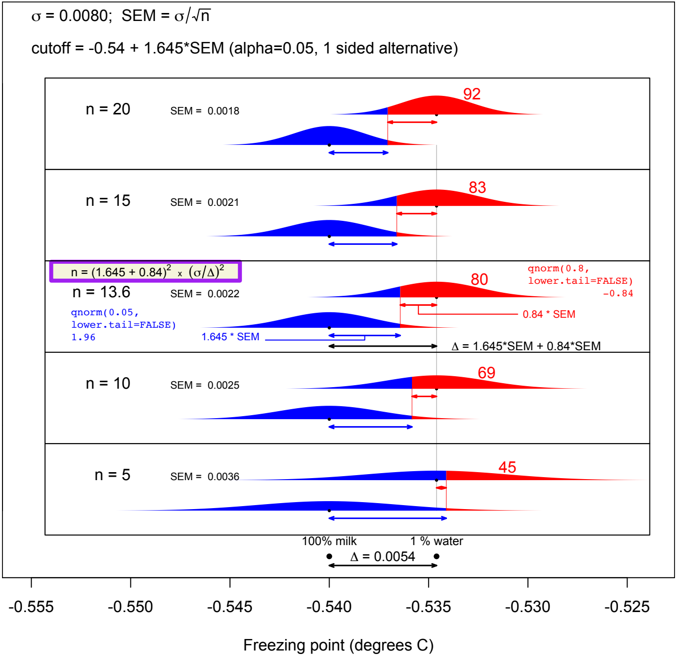


When the buyer uses samples of size 10



When the buyer uses samples of size 15




σ = 0.0080; SEM = σ/√n
cutoff = -0.54 + 1.645*SEM (alpha=0.05, 1 sided alternative)
n = 20
SEM = 0.0018
92
n = 15
SEM = 0.0021
83
n = (1.645 + 0.84)2 x (σ/Δ)2
n = 13.6
SEM = 0.0022
80
qnorm(0.8, lower.tail=FALSE) -0.84
0.84 * SEM
qnorm(0.05, lower.tail=FALSE) 1.96
1.645 * SEM
Δ = 1.645*SEM + 0.84*SEM
n = 10
SEM = 0.0025
69
n = 5
SEM = 0.0036
45
100% milk
1 % water
Δ = 0.0054
-0.555
-0.550
-0.545
-0.540
-0.535
-0.530
-0.525
Freezing point (degrees C)

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

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.

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:

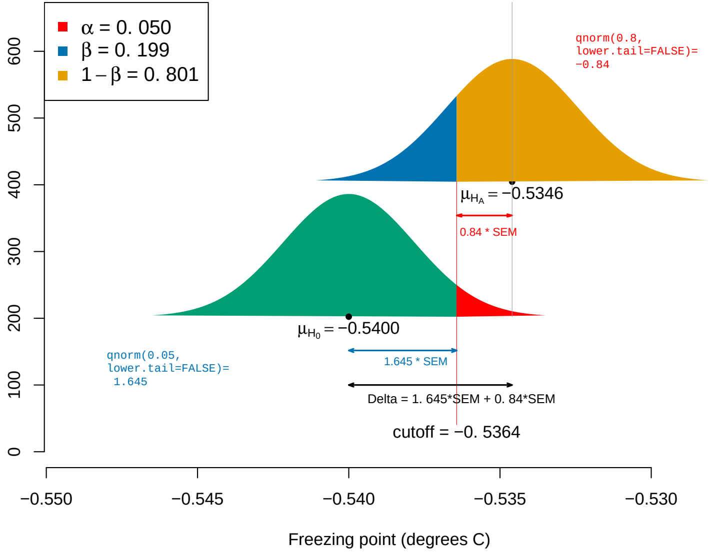
```
  stats::pnorm(cutoff, mean = mu.mixture, sd =
sigma/sqrt(n), lower.tail=FALSE)
```

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

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,

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