In-class exercise - Inference for means and Power Calculations. Solutions.

EPIB607 - Inferential Statistics[a]

[a] Fall 2019, McGill University

This version was compiled on October 10, 2019

In this exercise you will practice calculating confidence intervals using the t-distribution and the bootstrap.

Sampling distribution | Standard error | Normal distribution | Quantiles | Percentiles | Z-scores

R Code	Value
`qnorm(p = c(0.05, 0.95))`	-1.64, 1.64
`qnorm(p = c(0.025, 0.975))`	-1.96, 1.96
`qnorm(p = c(0.005, 0.995))`	-2.58, 2.58
`qt(p = c(0.025, 0.975), df = 400-1)`	-1.97, 1.97
`qt(p = c(0.025, 0.975), df = 25-1)`	-2.06, 2.06
`qt(p = c(0.025, 0.975), df = 20-1)`	-2.09, 2.09
`qt(p = c(0.025, 0.975), df = 16-1)`	-2.13, 2.13

1. Food intake and weight gain

If we increase our food intake, we generally gain weight. Nutrition scientists can calculate the amount of weight gain that would be associated with a given increase in calories. In one study, 16 nonobese adults, aged 25 to 36 years, were fed 1000 calories per day in excess of the calories needed to maintain a stable body weight. The subjects maintained this diet for 8 weeks, so they consumed a total of 56,000 extra calories. According to theory, 3500 extra calories will translate into a weight gain of 1 pound. Therfore we expect each of these subjects to gain 56,000/3500=16 pounds (lb). Here are the weights (given in the `weightgain.csv` file) before and after the 8-week period expressed in kilograms (kg):

```r
weight <- read.csv("weightgain.csv")
```

a. Calculate a 95% confidence interval for the mean weight change and give a sentence explaining the meaning of the 95%. State your assumptions.

```r
weight <- read.csv("~/git_repositories/EPIB607/exercises/inferencemeans/weightgain.csv")

# Creating new variable for weight change
weight$change <- weight$after-weight$before
weight$change_lb <- weight$change*2.2

# Calculating the mean of weight change and rounding
(ybar_change <- mean(weight$change))
```

```
#  [1] 4.73125
```

```
# Calculating the sample standard deviation
(ssd_change <- sd(weight$change))
```

```
#  [1] 1.745745
```

```
# sample size
(n <- nrow(weight))
```

```
#  [1] 16
```

```
# Calculating a 95% confidence interval version 1
qt_scaled <- function(p, df, mean, sd) {
  mean  + qt(p = p, df = df) * sd
}

(q1_ci95 <- qt_scaled(p = c(0.025, 0.975),
                      df = nrow(weight) - 1,
                      mean = ybar_change,
                      sd = ssd_change / sqrt(n)))
```

```
#  [1] 3.801008 5.661492
```

```
# Calculating a 95% confidence interval version 2
ybar_change +  qt(p = c(0.025, 0.975), df = n - 1) * ssd_change / sqrt(n)
```

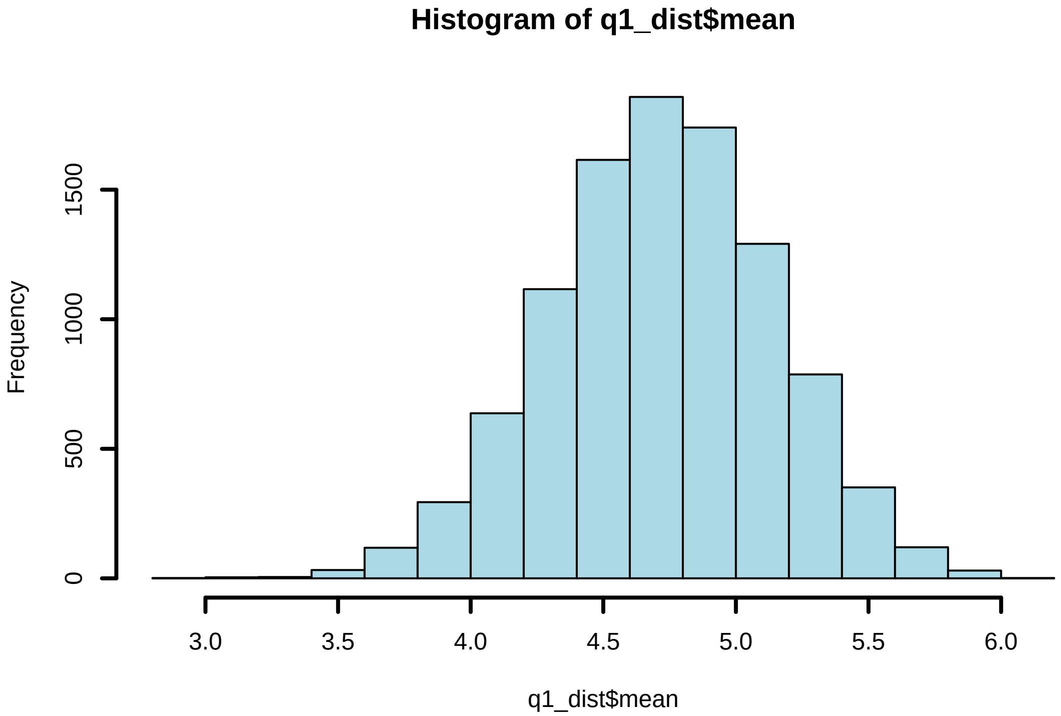
```
#  [1] 3.801008 5.661492
```

The 95% confidence interval for the mean weight change is 4.73 kg (3.8 kg, 5.66 kg). If the method used in this study were repeated many times, 95% of the time, the interval 3.8 kg and 5.66 kg will cover the true mean weight change. We can also say that we are 95% confident that the population mean weight gain is between (3.8 kg, 5.66 kg). Remember that the our uncertainty is about whether the particular sample we have at hand is one of the successful ones or one of the 5% that fail to produce an interval that captures the true value.

As this confidence interval was calculated using the *t* procedure, we are assuming that (1) we can regard our data as a simple random sample (SRS) from the population, (2) we have a representative sample of the population weight change and (3) observations of weight change in the population have a Normal distribution because we don't believe the CLT has kicked in.

b. Calculate a 95% bootstrap confidence interval for the mean weight change and compare it to the one obtained in part (a). Comment on the bootstrap sampling distribution and compare it to the assumptions you made in part (a).

```
q1_dist <- do(10000) * mean( ~ change, data = resample(weight))
hist(q1_dist$mean, col = "lightblue", lwd = 2)
```



```
round(quantile(~ mean, data = q1_dist, probs = c(0.001, 0.005, .025, 0.05,
                                                 0.90, 0.95, 0.975, 0.99)),2)
```

```
#   0.1%  0.5%  2.5%    5%   90%   95% 97.5%   99%
#   3.42  3.62  3.88  4.02  5.26  5.41  5.53  5.67
```

The 95% Bootstrap interval is given by [3.88, 5.53] kg. Very similar to the interval given in part a). Gives us some more confidence that the CLT has indeed kicked in.

c. Convert the units of the mean weight gain and 95% confidence interval to pounds. Note that 1 kilogram is equal to 2.2 pounds. Test the null hypothesis that the mean weight gain is 16 lbs. State your assumptions and justify your choice of test. Be sure to specify the null and alternative hypotheses. What do you conclude? We convert the Bootsrap CI by simply multiplying the upper and lower limits by 2.2lbs to give:

```
quantile(~ mean, data = q1_dist, probs = c(.025, 0.975)) * 2.2
```

```
#      2.5%    97.5%
#   8.53875 12.15500
```

The null hypothesis is that the theory of weight gain is the same as the measured weigth gain, $H_0 : \mu = \mu_o = 16$ lbs, and the alternative hypothesis is $H_a : \mu \neq 16$ lbs (two tailed test). I want to test it using the bootstrap method beacause I don't want to assume that the CLT has kicked in and the sampling distribution is normal. Since the upper limit of the confidence interval is below 16 lbs, there is evidence to suggest that we should reject the null hypothesis, i.e., the actual weight gain might be lower than the theory says.

2. Attitudes toward school

The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures the motivation, attitude toward school, and study habits of students. Scores range from 0 to 200. The mean score for U.S. college students is about 115, and the standard deviation is about 30. A teacher who suspects that older students have better attitudes toward school gives the SSHA to 25 students who are at least 30 years of age. Their mean score is $\bar{y} = 132.2$ with a sample standard deviation $s = 28$.

a. The teacher asks you to carry out a formal statistical test for her hypothesis. Perform a test, provide a 95% confidence interval and state your conclusion clearly.

2.1. Using the t procedure.

```
qt_scaled(p = c(0.025, 0.975), df = 24, mean = 132.2, sd = 28/sqrt(25))

# alternatively
132.2 + qt(p = c(0.025, 0.975), df = 24) * 28 / sqrt(25)
```

The null hypothesis is that older student mean SSHA score equals all student mean SSHA scores, $H_0 : \mu = \mu_0 = 115$. The alternate hypothesis is that older student mean SSHA score is greater than 115, $H_a : \mu > 115$ (ie: one-tail test). The sample size is on the smaller side (ie: below 30), so I chose to use a one sample t-test because I don't trust the sample sd to be a good estimate of the population sd. A two tailed 95% CI is [120.64, 143.76]. Based on this, I reject the null hypothesis, .i.e. , our data provides evidence that there might be a difference in SSHA scores between older students and the general population of students.

2.2. Using the z procedure.

```
qnorm(p = c(0.025, 0.975), mean = 132.2, sd = 30/sqrt(25))
```

```
#  [1] 120.4402 143.9598
```

```
# alternatively
132.2 + qnorm(p = c(0.025, 0.975)) * 30/sqrt(25)
```

```
#  [1] 120.4402 143.9598
```

Assuming that the standard deviation of the all U.S. college students of 30 is accurate and taken as sigma, the 95% confidence interval for the population mean SSHA score is [120.44, 143.96].

b. What assumptions did you use in part (a). Which of these assumptions is most important to the validity of your conclusion in part (a).

We are assuming that this is a simple random sample of older students. If using the t-distribution, we are assuming that the standard deviation of the population is not a good estimate of the standard deviation of our sample. The most important assumption we've made is that the CLT has kicked in and therefore the sampling distribution is normal.

If using z procedure → that the standard deviation of the all U.S. college students of 30 is accurate and taken as sigma, and that the sample size is enough that the CLT has kicked in.

3. Does a full moon affect behavior?

Many people believe that the moon influences the actions of some individuals. A study of dementia patients in nursing homes recorded various types of disruptive behaviors every day for 12 weeks. Days were classified as moon days if they were in a 3-day period centered at the day of the full moon. For each patient, the average number of disruptive behaviors was computed for moon days and for all other days. The hypothesis is that moon days will lead to more disruptive behavior. We look at a data set consisting of observations on 15 dementia patients in nursing homes (available in the `fullmoon.csv` file):

```
fullmoon <- read.csv("fullmoon.csv")
```

```
#     patient moon_days other_days
#  1        1      3.33       0.27
#  2        2      3.67       0.59
#  3        3      2.67       0.32
#  4        4      3.33       0.19
#  5        5      3.33       1.26
#  6        6      3.67       0.11
#  7        7      4.67       0.30
#  8        8      2.67       0.40
#  9        9      6.00       1.59
#  10      10      4.33       0.60
#  11      11      3.33       0.65
#  12      12      0.67       0.69
#  13      13      1.33       1.26
#  14      14      0.33       0.23
#  15      15      2.00       0.38
```

a. Calculate a 95% confidence interval for the mean difference in disruptive behaviors. State the assumptions you used to calculate this interval.

```
(moon.mean <- mean(moon.diff))
```

```
#  [1] 2.432667
```

```
(moon.sdev <- sd(moon.diff))
```

```
#  [1] 1.46032
```

```
(moon.n <- length(moon.diff))
```

```
#  [1] 15
```

```
qt_scaled(p=c(0.025, 0.975), df = moon.n - 1, mean = moon.mean, sd = moon.sdev/sqrt(moon.n))
```

```
#  [1] 1.623968 3.241365
```

```
# alternatively
moon.mean + qt(p = c(0.025, 0.975), df = moon.n - 1) * moon.sdev/sqrt(moon.n)
```

```
#  [1] 1.623968 3.241365
```

Assuming a simple random sample and that our sample is representative of the population, and that the difference in distruptive events is normally distributed in the population (OR you can say that you believe the Central Limit Theorem (CLT) has kicked in), I calculated a 95% confidence interval for the population mean difference of [1.62, 3.24 distruptive events on full moon days when compared to other days. This was done using the t procedure due to the unknown sigma and small sample size of n = 15.

b. Test the hypothesis that moon days will lead to more disruptive behavior. State your assumptions and provide a brief conclusion based on your analysis.

```
#t-statistic
(t_statistic <- (moon.mean - 0) / (moon.sdev/sqrt(moon.n)))
```

```
#  [1] 6.451789
```

```
# p-value
pt(q = t_statistic, df = moon.n-1, lower.tail = F)
```

```
#  [1] 7.590761e-06
```

$H_0 : \mu = 0$ (i.e. that there is no difference in distruptive behaviours on full moon nights when compared to other nights) and $H_a : \mu > 0$ (i.e. that there are more distruptive behaviours on full moon nights when compared to other nights).

I calculate a one-sided (to the right) p-value using the t statistic (because unknown sigma and small sample size of n = 15) of 7.5907605×10^{-6}. Assuming the null hypothesis is true, there is very small probability that the observed mean difference of 2.4326667 came from the null distribution. We could also say that since $\mu = 0$ is not contained within our 95% confidence interval, this provides evidence against the null hypothesis. This is based on the assumptions that the sample distribution here is enough that the CLT has kicked in as well.

c. Find the minimum value of the mean difference in disruptive behaviors ($\bar{y}$) needed to reject the null hypothesis.

We first need to figure out for what values we can reject the null. Since this is a one-sided alternative we want and $\alpha = 0.05$ in the right tail. The cutoff is then given by:

```
(tscore.null <- qt(p = 0.95, df = 15-1))
```

```
#  [1] 1.76131
```

Then we solve for $\bar{y}$ in the t-statistic formula:

$$
\begin{aligned}
t_{statistic} &= \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \\
1.76 &= \frac{\bar{y} - 0}{1.46/\sqrt{15}}
\end{aligned}
$$

The solution to the above equation is:

```
tscore.null*(moon.sdev/sqrt(moon.n))
```

```
#  [1] 0.6641074
```

Assuming H_0 is true, under the null distribution the minimum value in mean difference in distruptive behaviours needed to reject the null hypothesis at an $\alpha = 0.05$ level is 0.66 distruptive behaviours.

d. What is the probability of detecting an increase of 1.0 aggressive behavior per day during moon days?

$$
H_0 : \mu = 0 \qquad\qquad H_A : \mu > 1
$$

This is a standard power calculation. In the first step, we calculate the difference in mean disruptive behaviours needed to reject the null hypothesis, which we calculated in part c) to be 0.66. The corresponding t value under the alternative hypothesis distribution is given by

$$
\begin{aligned}
t_{statistic} &= \frac{\bar{y} - \mu_A}{s/\sqrt{n}} \\
&= \frac{0.664 - 1}{1.46/\sqrt{15}} \\
&= -0.890
\end{aligned}
$$

We then calculate the probability of observing this value or more under the alternative hypothesis distribution:

```
pt(q = -0.890, df = moon.n - 1, lower.tail = FALSE)
```

```
#  [1] 0.805748
```

Therefore, the power to detect an increase of 1.0 aggressive behavior per day during moon days is 80.57%.

4. 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.

a) Does this sample provide evidence to reject the null hypothesis of no difference between children of Lake Wobegon and the general population?

$$H_0 : \mu = 100 \qquad\qquad H_A : \mu > 100$$

The p-value (one-sided test) is given by:

```r
pnorm(q = 112.8, mean = 100, sd = 15 / sqrt(9), lower.tail = FALSE)
```

```
#  [1] 0.005233608
```

This sample provides evidence against the null hypothesis. We calculated a p-value of 0.01 which tells us the probability of observing the sample mean of 112.8 under the null hypothesis distribution is very unlikely.

b) 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?

$$H_0 : \mu = 100 \qquad\qquad H_A : \mu > 110$$

Step 1 is to calculate the cutoff in order to reject the null. This is given by

```r
# this is a one-sided alternative so we want alpha=5% in the right tail
(cutoff <- qnorm(p = 0.95, mean = 100, sd = 15 / sqrt(9)))
```

```
#  [1] 108.2243
```

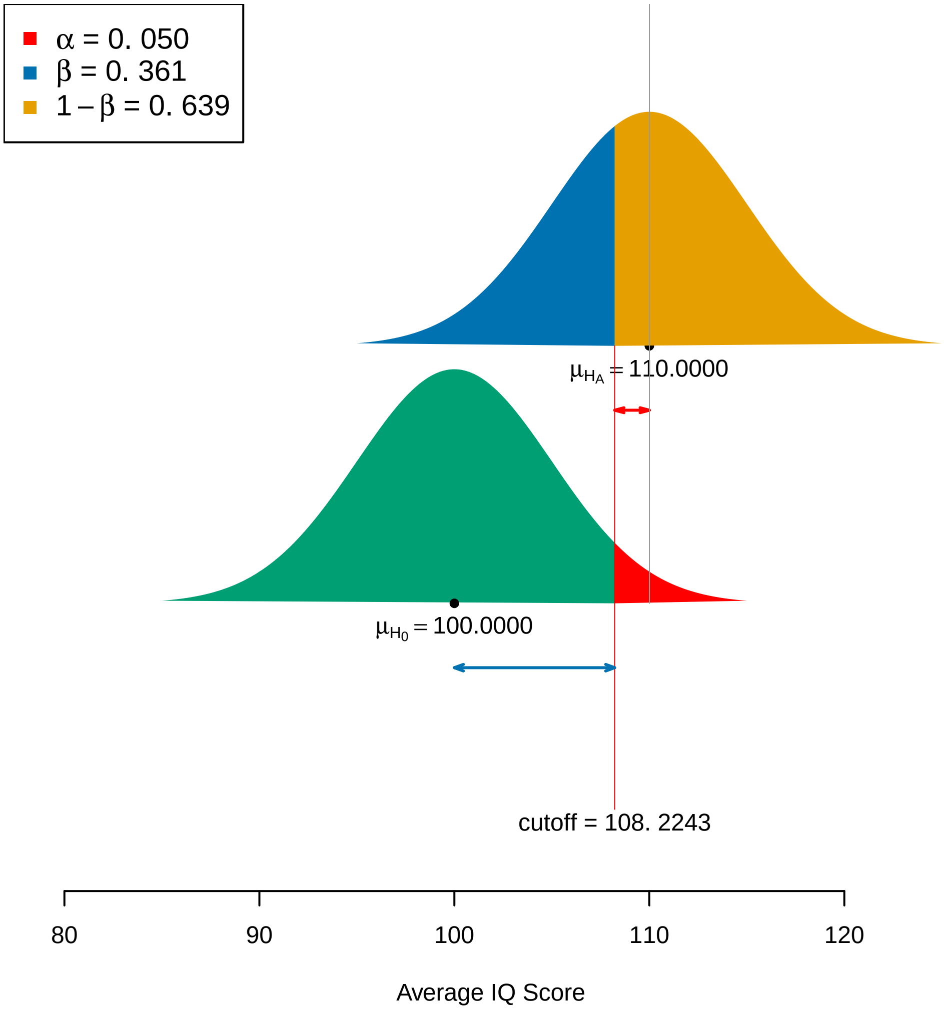
Then we calculate the probability of observing this cutoff or greater under the alternative hypothesis:

```r
pnorm(q = cutoff, mean = 110, sd = 15 / sqrt(9), lower.tail = FALSE)
```

```
#  [1] 0.63876
```

The following figure visualizes this calculation:

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



c) If you hoped to use a **two-sided** test to show that the children from Lake Wobegon are at least 5 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?

$$H_0 : \mu = 100 \qquad H_A : \mu = 105$$

Because its a two-sided test, we need to find the both cutoffs, i.e., the values of the sample mean that will reject the null. This is given by:

```
# two-sided test at alpha=5% means we want 2.5% in the tails
(cutoffs <- qnorm(c(0.025, 0.975), 100, 15 / sqrt(9)))
```

```
#  [1]  90.20018 109.79982
```

That is, we will reject the null hypothesis if the sample mean is 90.2 or less, OR reject than null if the sample mean is 109.8 or more.

Next we need to calculate these probabilities under the alternative hypothesis:

```
# left tail probability
(p_left <- pnorm(q = 90.20018, mean = 105, sd = 15 / sqrt(9), lower.tail = TRUE))
```

```
#  [1] 0.001538375
```

```
# right tail probability
(p_right <- pnorm(q = 109.79982, mean = 105, sd = 15 / sqrt(9), lower.tail = FALSE))
```

```
#  [1] 0.1685367
```

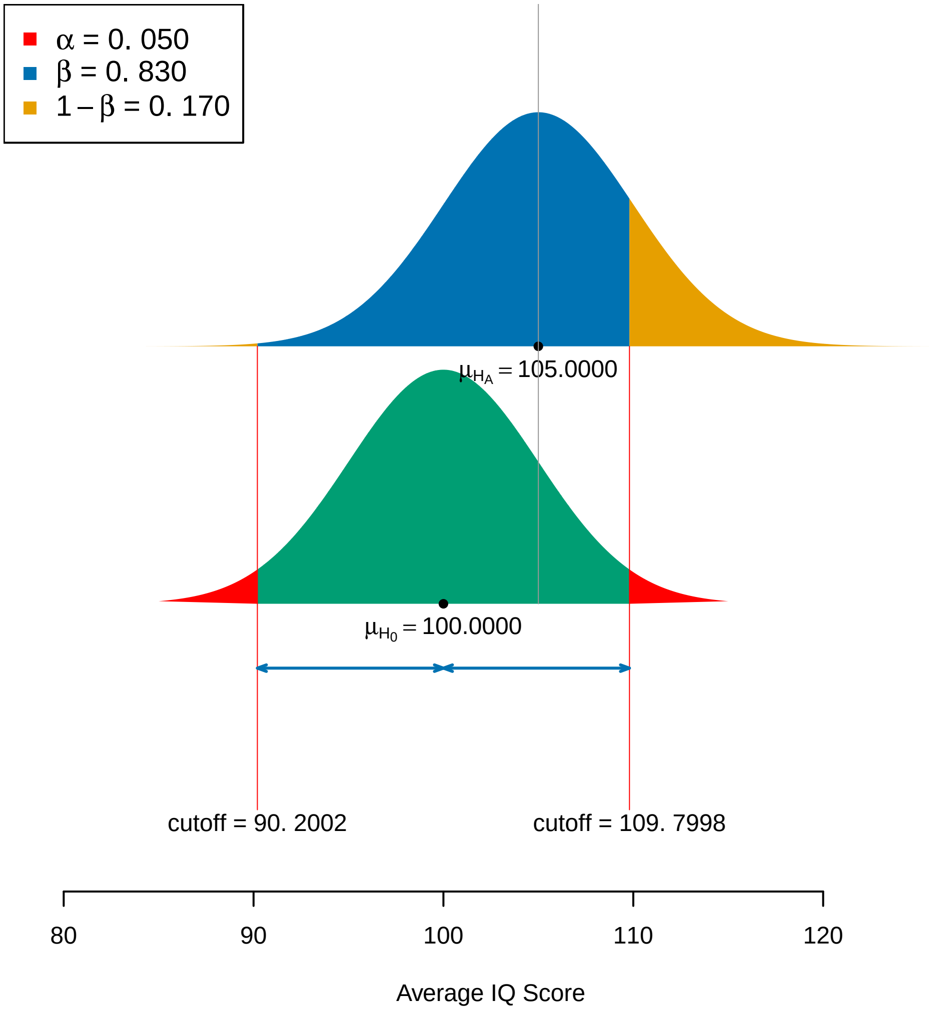
And the power is the sum of these two probabilities:

```
p_left + p_right
```

```
#  [1] 0.170075
```

As show in the following figure:

```
power_plot(n = 9, s = 15, mu0 = 100, mha = 105,
           cutoff = qnorm(c(0.025, 0.975), 100, 15 / sqrt(9)),
           alternative = "equal", xlab = "Average IQ Score")
```



5. Bias in step counters

Following the study by Case et al., JAMA, 2015, suppose we wished to assess, via a formal statistical test, whether (at an *population*, rather than an individual, level) a step-counting device or app is unbiased (H_0) or under-counts (H_A). Suppose we will do so the way Case et al. did, but measuring n persons just once each. We observe the device count when the true count on the treadmill reaches 500.

a. Using a planned sample size of $n = 25$, and $\sigma = 60$ steps as a pre-study best-guess as to the s that might be observed in them, calculate the critical value at $\alpha = 0.01$.

```
qnorm(p = 0.01, mean = 500, sd = 60/sqrt(25))
```

```
#  [1] 472.0838
```

The critical value, i.e., the mean step counts to reject the null is 472.08 steps.

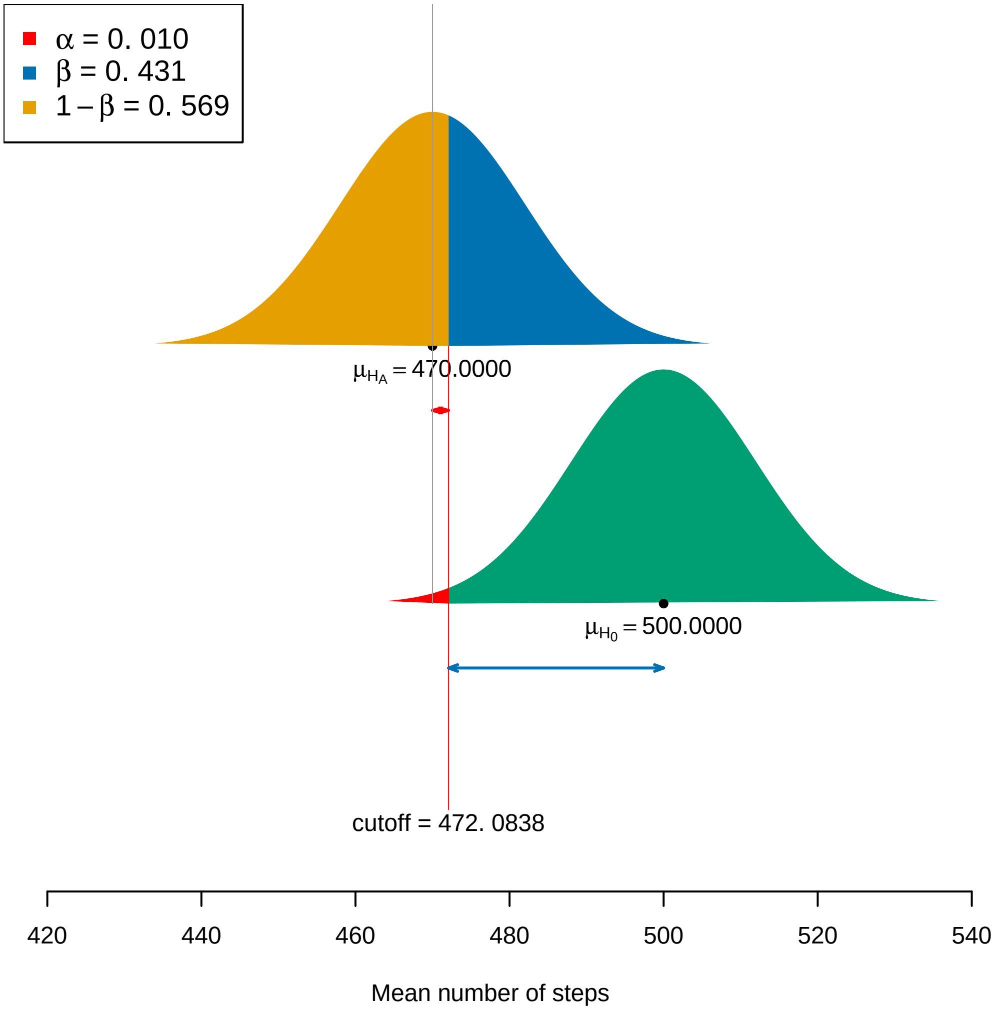
b. Now imagine that the mean would not be the null 500, but $\mu = 470$. Calculate the probability that the mean in the sample of 25 will be less than this critical value. Use the same σ for the alternative that you used for the null. What is this probability called?

```
critical_z <- qnorm(p = 0.01, mean = 500, sd = 60/sqrt(25))
pnorm(q = critical_z, mean = 470, sd = 60/sqrt(25), lower.tail = TRUE)
```

```
#  [1] 0.5689306
```

The probability of getting a sample of size 25 with a mean less than the critical value of 472.08 steps is 0.5689306. This is the statistical power to detect a mean difference of 30 fewer steps.

```
power_plot(n = 25, s = 60, mu0 = 500, mha = 470,
           cutoff = critical_z,
           alternative = "less", xlab = "Mean number of steps")
```



c. Determine the sample size required to detect a 30 step mean decrease in steps with 80% power using a 1% level of significance. Plot the null and alternative distributions in a diagram using the `plot_power` function.

Under the null hypothesis, we know that the mean step count has to be 2.32 standard errors of the mean away from $\mu_0 = 500$ in order to reject the null hypothesis at an $\alpha = 0.01$. 2.32 is the z value such that there is 0.01 area in the left tail of the null distribution:

```
qnorm(p = 0.01, lower.tail = TRUE)
```

```
#  [1] -2.326348
```

Under the alternative hypothesis, we know that the mean step count has to be 0.84 standard errors of the mean (SEM) away from $\mu_A = 470$ such that there is 20% area in the right tail of the alternative hypothesis distribution:

```
qnorm(p = 0.20, lower.tail = FALSE)
```

```
#  [1] 0.8416212
```

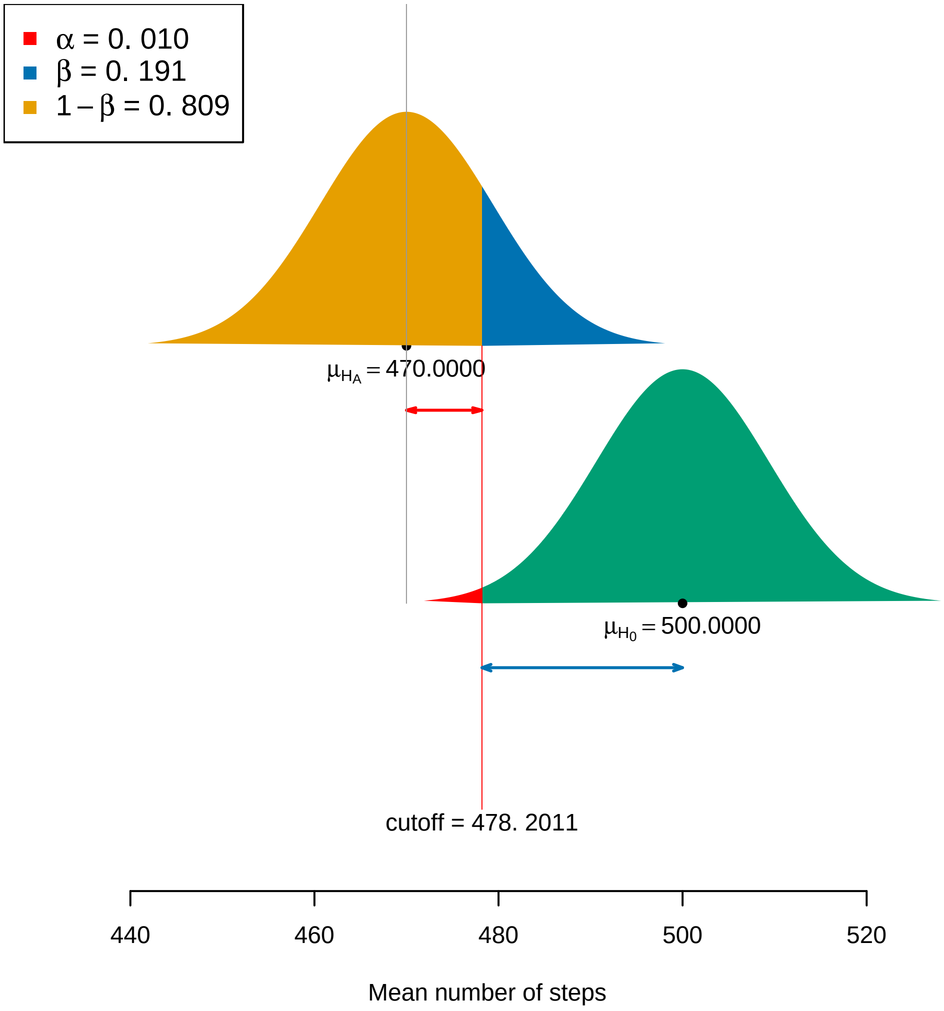
We know that the distance between $\mu_0 = 500$ and $\mu_A = 470$ must be equal to 0.84SEM + 2.32SEM. Note that although the quantile calculated under the null is negative, since we are dealing with distance, we use the absolute value. This is the balancing equation:

$$
\begin{aligned}
\Delta &= 0.84 \times SEM + 2.32 \times SEM \\
\Delta &= (0.84 + 2.32) SEM \\
&= (0.84 + 2.32) \frac{\sigma}{\sqrt{n}} \\
\sqrt{n} &= (0.84 + 2.32) \frac{\sigma}{\Delta} \\
n &= (0.84 + 2.32)^2 \left(\frac{\sigma}{\Delta} \right)^2 \\
&= (0.84 + 2.32)^2 \left(\frac{60}{30} \right)^2 \\
&= 40.14
\end{aligned}
$$

Therefore we need 41 subjects to detect a 30 step mean decrease in steps with 80% power using a 1% level of significance.

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

power_plot(n = 41,
           s = 60,
           mu0 = 500,
           mha = 470,
           cutoff = qnorm(0.01, mean = 500, sd= 60/sqrt(41), lower.tail = TRUE),
           alternative = "less",
           xlab = "Mean number of steps ")
```


α = 0. 010
β = 0. 191
1 – β = 0. 809
$\mu_{H_A} = 470.0000$
$\mu_{H_0} = 500.0000$
cutoff = 478. 2011
440
460
480
500
520
Mean number of steps