

Lecture 5: Statistical power

Matti Pirinen

20.9.2021

Intro to statistical power

Let's return to the setting of Lectures 2 & 3: n pairs of same-sex twins discordant for psoriasis and in x of the pairs the psoriatic individual has larger BMI than the non-psoriatic one.

Let's see how the sample size n affects our inference on the proportion p of twin pairs where psoriatic twin has higher BMI than the non-psoriatic one. Let's consider two data sets with sample sizes of 20 and 100, and let's assume that the point estimate for the proportion in both cases is 70%. (That is, we have observed 14 and 70 successes from data sets of sizes 20 and 100, respectively.) Let's do the binomial test in both data sets.

```
p.est = 0.70 #our point estimate, x/n, is 70% in both cases
n = c(20, 100) #two sample sizes
x = n*p.est #two observations, one for each sample size, both correspond to proportion of 70%
x #14 and 70
```

```
## [1] 14 70
```

```
binom.test(x[1], n[1])
```

```
##
## Exact binomial test
##
## data: x[1] and n[1]
## number of successes = 14, number of trials = 20, p-value = 0.1153
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4572108 0.8810684
## sample estimates:
## probability of success
##                0.7
```

```
binom.test(x[2], n[2])
```

```
##
## Exact binomial test
##
## data: x[2] and n[2]
## number of successes = 70, number of trials = 100, p-value = 7.85e-05
## alternative hypothesis: true probability of success is not equal to 0.5
```

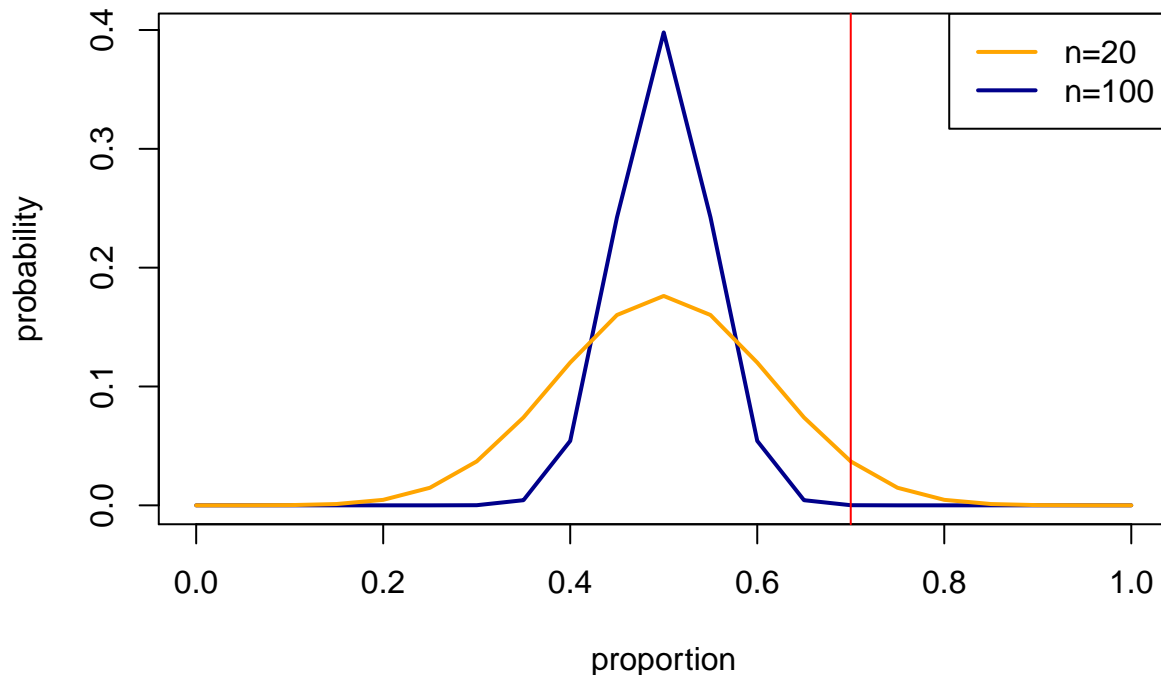
```
## 95 percent confidence interval:
## 0.6001853 0.7875936
## sample estimates:
## probability of success
## 0.7
```

We see that the confidence interval around the point estimate 0.70 shrinks as the sample size n grows. This is because with larger n we have more information about where exactly the proportion parameter p is, and this is reflected by a smaller standard error and narrower CI. With more information there is less uncertainty about p . This is why statisticians always want to have bigger sample size.

There is also a striking difference in P-values (0.12 vs $8e-5$). Let's try to understand why the sample size affects statistical testing and P-value.

Null hypothesis is that BMI and psoriasis are independent. This is equivalent to the assumption that the observed value follows $\text{Bin}(n, p = 0.5)$ distribution, that is, it is like the total number of heads in n tosses of a fair coin. Let's plot in the same figure the probability mass functions under the null hypothesis for possible point estimate values (of the proportion $p = x/n$) for the two sample sizes.

```
p.range = seq(0, 1, 0.05) #values at which density will be evaluated at steps of 1/20=0.05
n = c(20, 100)
y.1 = dbinom(p.range*n[1], size = n[1], p = 0.5) #null distribution when n = 20
y.2 = dbinom(p.range*n[2], size = n[2], p = 0.5) #null distribution when n = 100
#renormalize so that both distributions sum to 1 and are thus comparable
y.1 = y.1 / sum(y.1)
y.2 = y.2 / sum(y.2)
plot(p.range, y.2, col = "darkblue", t = "l", lwd = 2, xlab = "proportion", ylab = "probability")
lines(p.range, y.1, col = "orange", t = "l", lwd = 2) #add line for n=20
legend("topright", legend = paste0("n=",n), col = c("orange","darkblue"), lty = 1, lwd = 2)
abline(v = 0.7, col = "red") # Mark observation 70% in red
```



We see that the null distribution for larger sample size $n=100$ is more concentrated near the null hypothesis value 0.5 than the distribution with $n=20$. By looking at the probability mass outside the red line, we can

conclude that if the true association between BMI and psoriasis was such that in 70% of discordant twin pairs the psoriatic one had a larger BMI, then, with sample size $n = 100$, we would expect to distinguish that association from the null hypothesis with clear statistical evidence (small P-value because observation is so improbable under the blue null), but, for $n = 20$, we might not distinguish it clearly (P-value can well be > 0.05 because the observation near 70% could happen also under the orange null every now and then). We say that the larger sample size brings larger **statistical power** to detect true deviations from the null hypothesis.

This logic is verified by the binomial tests above, where we saw that P-values for observation 70% in these two sample sizes are quite different: 0.12 for $n = 20$ but less than $1e-4$ for $n = 100$.

Example 5.1. Suppose that the biomarker B is distributed like $N(1,1)$ in healthy population and like $N(2,1)$ in diabetes cases. Generate a random set of n samples from the healthy population and another sample of same size from diabetics. We know that the distributions have different means (namely, 1 and 2), and we want to see how statistically clearly we could detect that difference with different sample sizes. Test whether the means are different using a t-test when

- (i) $n = 5$
- (ii) $n = 25$
- (iii) $n = 50$

Do the means look statistically different based on the P-values in these three cases?

```
ns = c(5, 25, 50)
for(n in ns){ #With for loop we do not need to copy-paste the code separately for each value
  x.1 = rnorm(n, 1, 1) #healthy
  x.2 = rnorm(n, 2, 1) #diabetics
  print(t.test(x.1, x.2)) #Needs "print()" inside for-loop to print out the result of t-test.
}
```

```
##
## Welch Two Sample t-test
##
## data: x.1 and x.2
## t = -2.2419, df = 7.0118, p-value = 0.05985
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -2.09460386 0.05545062
## sample estimates:
## mean of x mean of y
## 0.857512 1.877089
##
##
## Welch Two Sample t-test
##
## data: x.1 and x.2
## t = -2.3704, df = 43.329, p-value = 0.02229
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1.3441528 -0.1085158
## sample estimates:
## mean of x mean of y
```

```
## 1.297058 2.023392
##
##
## Welch Two Sample t-test
##
## data: x.1 and x.2
## t = -5.4251, df = 97.976, p-value = 4.185e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1.522608 -0.707022
## sample estimates:
## mean of x mean of y
## 1.041062 2.155877
```

P-values get smaller with increasing sample size, reflecting that we will detect the same true difference in means with more statistical evidence using a larger sample size than using a smaller sample size.

P-values and Type I error rate

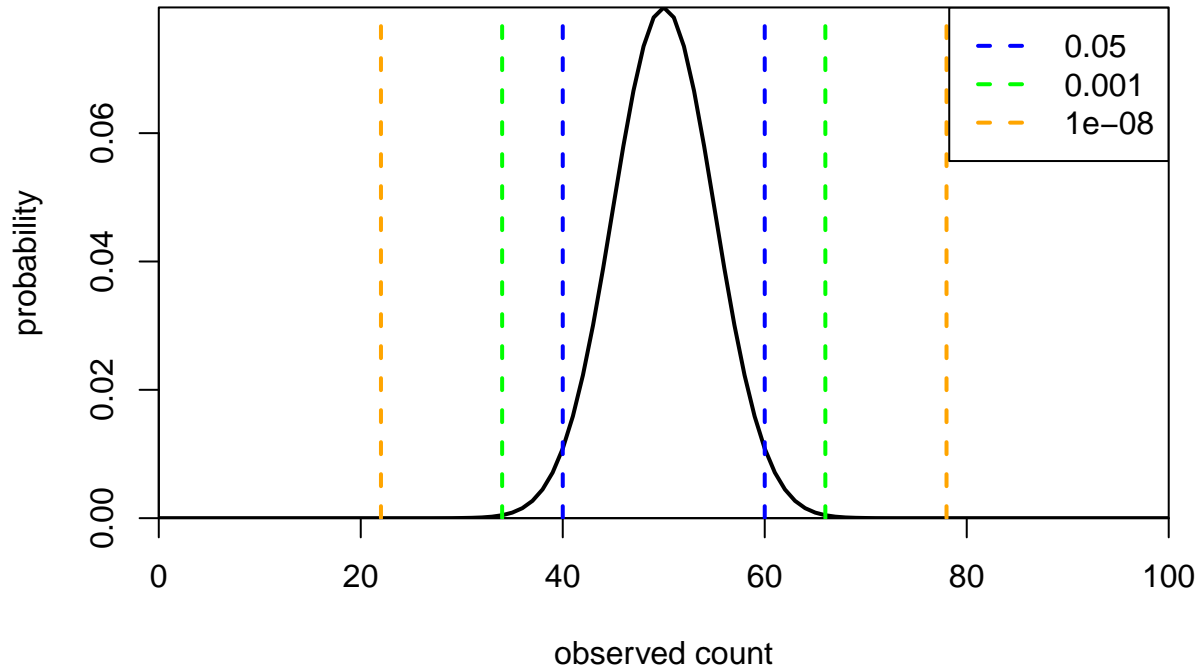
So far we have observed data and computed the P-value of observing such kind of data under the null hypothesis. The idea has been that if P-value is small, then the observed data seem unlikely to arise under the null hypothesis and hence we have a reason to suspect the null hypothesis.

We can also imagine a statistical testing procedure where before seeing data and computing P-value we set a significance level (α , “alpha”, e.g. 0.05, 0.01 or 0.001) and if the observed data reach that significance level (i.e. computed P-value $\leq \alpha$), we “reject” the null hypothesis because it seems implausible. The motivation for pre-defining the significance level is that then we know that the testing procedure has the following property: If the null hypothesis was true, then in repeated experiments we would falsely reject the null hypothesis for proportion α of the experiments. By controlling α , we can control the level of false rejections of the null hypothesis. If $\alpha = 0.05$, then we would make a false call in 1/20 experiments where the null hypothesis was actually true; if α is 0.001, then 1/1000 of the true null hypotheses were falsely rejected etc. The significance level α of the testing procedure is also called Type I error rate, since a false rejection of the null hypothesis is called a type I error. “**Type I error**: Falsely declaring an effect found when there truly is none.”

Let’s draw a picture of null distribution of 100 flips of a fair coin, (analogous to twin pairs where higher BMI and psoriasis status go together), and mark the 2-sided significance thresholds for 0.05, 0.001 and 1e-8.

```
n = 100 # twin pairs
p0 = 0.5 # null hypothesis value for proportion
x = 0:n # plotting range 0,...,n
#Plot the null distribution
plot(x, dbinom(x, n, p0), t = "l", lwd = 2, main = "null distribution Bin(100,0.5)",
     xlab = "observed count", ylab = "probability", xaxs = "i", yaxs = "i")
#Plot cutoff point for P-value of 0.001 under the null distribution in red
sig.threshold = c(0.05, 0.001, 1e-8)
cols = c("blue", "green", "orange")
for(ii in 1:length(sig.threshold) ){ #goes through all sig.thresholds
  abline(v = qbinom(sig.threshold[ii]/2, n, p0),
        col = cols[ii], lwd = 2, lty = 2) #mark left cutpoint
  abline(v = qbinom(1-sig.threshold[ii]/2, n, p0),
        col = cols[ii], lwd = 2, lty = 2) # mark also right cutpoint
}
legend("topright", col = cols, legend = sig.threshold, lwd = 2, lty = 2)
```

null distribution Bin(100,0.5)



We see that in order to get a significant deviation from the null hypothesis at significance level 0.05, we need to have either <40 or >60 successes. On the other hand, to get statistical significance at level $1e-8$, we would need as extreme a result as <22 or >78 successes. This reflects the fact that we need much more evidence against the null in order to reject the null hypothesis at level $1e-8$ than we need to reject the null at level 0.05. Consequently, we will make less Type I errors if our significance level is $1e-8$ than if it is 0.05. But at a same time, at level $1e-8$, we will miss some real deviations from the null hypothesis that we would detect at level 0.05. That is, we do more Type II errors at a lower significance level than at a higher significance level. “**Type II error:** Not rejecting the null hypothesis when it does not hold.”

Typically significance level is chosen purely by Type I error rate, that is, by our tolerance for false rejection of null hypothesis. This tolerance varies between contexts. In clinical trials, the significance level could be 0.05, but for example in genome-wide analyses we get excited about a new association between a genetic variant and a disease only when it reaches a significance level $5e-8 = 0.00000005$! Similarly, in 2012 the physicists claimed that they had “found” Higgs’ boson after they had gathered statistical evidence that reached their “5 sigma rule” (observation 5 standard deviations away from the null hypothesis value) that corresponds to a significance level of $6e-7$.

Rationale for differences could be that in clinical trial we only study one hypothesis (a therapy works) that possibly already has some prior evidence why it has been developed in the first place. So we do not require a lot statistical evidence at this point to get further with the process. In genome-wide analyses we are starting by a fishing experiment where we have millions of genetic variants and no prior evidence of any real association with diseases. Thus, we need a lot of statistical evidence to start believing that this one variant (out of millions of possible ones) is truly associated with the disease. Finally, with Higgs’ boson, we again have only one hypothesis to test, but here the price of making a wrong claim is large for the science: Physicists don’t want to publish a result about existence of a new particle unless they are (very) certain. Hence, they require much more statistical evidence than clinical trials.

Examples 5.2.

1. Your null hypothesis is that a coin is fair ($p=50\%$). If you toss the coin n times and count the number

of heads, what are the critical points below and above the null value 50% at which the outcome starts showing statistically significantly that the coin is biased, when

- (i) $n = 10, \alpha = 0.05$
- (ii) $n = 100, \alpha = 0.05$
- (iii) $n = 10, \alpha = 0.0001$
- (iv) $n = 100, \alpha = 0.0001$.

```
p = 0.5
ns = rep(c(10, 100), 2)
a = rep(c(0.05, 0.0001), each = 2)
for(ii in 1:length(ns)){ #for-loop over 4 sets of parameters n and alpha
  print(paste("n =",ns[ii],"alpha =",a[ii],": at most",
             qbinom(a[ii]/2,ns[ii],p)-1,"or at least",
             qbinom(1-a[ii]/2,ns[ii],p)+1))
}
```

```
## [1] "n = 10 alpha = 0.05 : at most 1 or at least 9"
## [1] "n = 100 alpha = 0.05 : at most 39 or at least 61"
## [1] "n = 10 alpha = 1e-04 : at most -1 or at least 11"
## [1] "n = 100 alpha = 1e-04 : at most 30 or at least 70"
```

We see that for $n = 10$, no observation can reach significance level $\alpha = 0.001$ as the critical points are outside the range $0, \dots, 10$.

2. Compute these critical points for a fixed $\alpha = 0.05$ as n goes through 10, 50, 100, 200, ... 1000. Compute which proportions they correspond to and show the results on screen.

```
a = 0.05
ns = c(10, 50, seq(100,1000,100))
res = c() #empty vector that will be filled to be a matrix with 3 columns:
#1st col = n; 2nd col = lower critical point; 3rd col = upper critical point
for(ii in 1:length(ns)){
  res = rbind(res,
             c(ns[ii], qbinom(a/2, ns[ii], p)-1, qbinom(1-a/2, ns[ii], p)+1))
}
res[,2:3] = res[,2:3]/res[,1] #proportions are columns 2 and 3 divided by column 1
res #show results on screen
```

```
##      [,1]      [,2]      [,3]
## [1,]   10 0.1000000 0.9000000
## [2,]   50 0.3400000 0.6600000
## [3,]  100 0.3900000 0.6100000
## [4,]  200 0.4250000 0.5750000
## [5,]  300 0.4400000 0.5600000
## [6,]  400 0.4475000 0.5525000
## [7,]  500 0.4540000 0.5460000
## [8,]  600 0.4583333 0.5416667
## [9,]  700 0.4614286 0.5385714
## [10,] 800 0.4637500 0.5362500
## [11,] 900 0.4666667 0.5333333
## [12,] 1000 0.4680000 0.5320000
```

We see that in order to get a significant result at significance level 0.05, with sample size of 10 we need <20% or >80% success rate, while with sample size of 1000 it is enough if we get <47% or >53% success rate. Sample size really matters when it comes to determining which point estimate value is significant at a fixed significance level!

Statistical Power

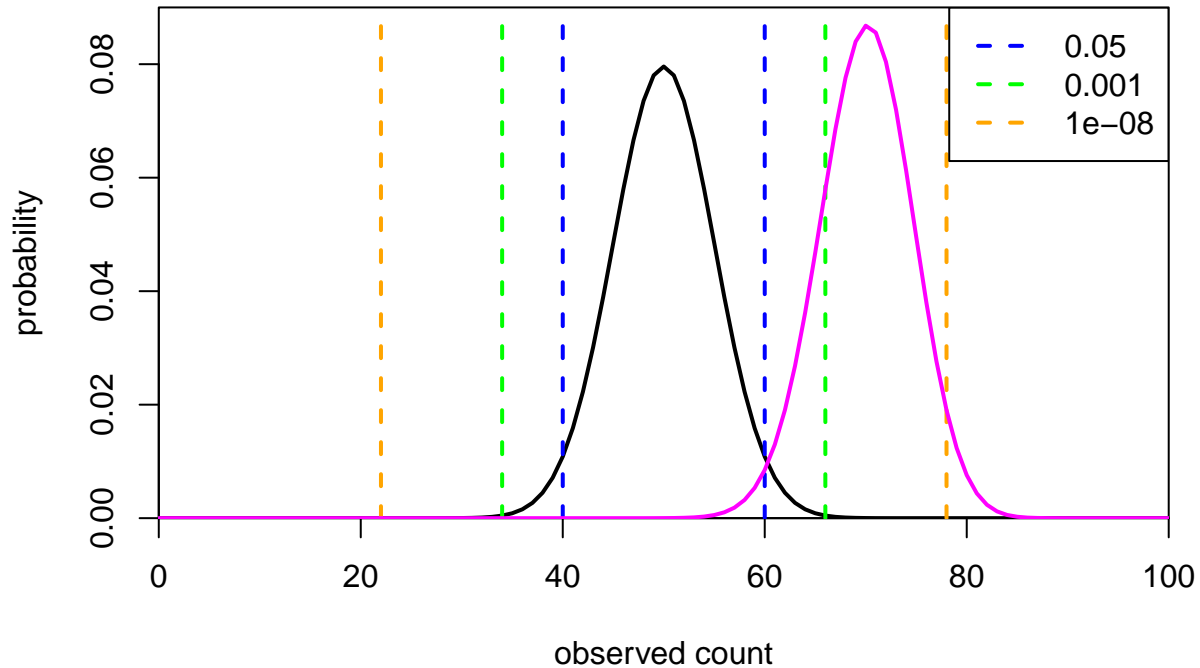
Statistical power is a probability of detecting a given effect size, with a given sample size at a given significance level.

Typical context where statistical power comes up is the following. Let's say that in order to be convinced that there is an association between psoriasis and BMI we want to see a P-value of 0.001 (or less) if true effect size is $p = 0.70$ (or larger), when we compare the observed twin-pair data to the null hypothesis value of $p = 0.50$. How large a study should we have in order that the power with these parameters is at least 90%?

Before computing the power, let's add to our picture of the significance regions also the distribution of the **alternative hypothesis**, that is defined by the true effect size parameter. Here the alternative hypothesis is $\text{Bin}(100, 0.7)$.

```
n = 100 #twin pairs
p0 = 0.5 #null hypothesis proportion
x = 0:n #plotting range 0,...,n
#Plot null distribution
plot(x, dbinom(x, n, p0), t = "l", lwd = 2, main = "null Bin(100,0.5); altern Bin(100,0.7)",
      xlab = "observed count", ylab = "probability", xaxs = "i", yaxs = "i", ylim = c(0,0.09))
#Plot cutoff point for P-value of 0.001 under the null distribution in red
sig.threshold = c(0.05, 0.001, 1e-8)
cols = c("blue", "green", "orange")
for(ii in 1:length(sig.threshold)){
  abline(v = qbinom(sig.threshold[ii]/2, n, p0),
         col = cols[ii], lwd = 2, lty = 2) #add left cutpoint
  abline(v = qbinom(1-sig.threshold[ii]/2, n, p0),
         col=cols[ii],lwd=2,lty=2) #add also right cutpoint
}
legend("topright", col = cols, legend = sig.threshold, lwd = 2, lty = 2)
#And add alternative hypothesis distribution in magenta
p1 = 0.7
lines(x, dbinom(x, n, p1), lwd = 2, col = "magenta")
```

null Bin(100,0.5); altern Bin(100,0.7)



Let's illustrate how the power calculation happens. For a fixed sample size n ($=100$), significance level α ($=0.001$) and effect size (null $p = 0.50$, alternative $p = 0.70$) we need to

1. Find the set S of possible observations that would give a P-value smaller than α under the null hypothesis. (Here this set consists of values outside the green lines.)
2. Find which proportion of the observations under the alternative hypothesis falls in set S (and hence would give a P-value $< \alpha$ under the null).

```
n = 100
a = 0.001
#Cutpoints for significant results at level 0.001 under the null value p = 0.5
left.cut = qbinom(a/2, n, 0.5)-1
right.cut = qbinom(1-a/2, n, 0.5)+1
#For observations 0...left.cut and right.cut...100 we would thus have P<0.001 for n=100 and p=0.5.
S = c(0:left.cut, right.cut:100) #Set of observations which give small enough P-value

#What is a probability of observing a value in set S, under the alternative?
#Let's sum the probabilities over the set S:
sum(dbinom(S, size = n, p = 0.7))
```

```
## [1] 0.7792578
```

Thus we have a power of about 80%, (more accurately 78%), to discover a deviation from the null hypothesis (50% success rate), at significance level 0.001 (that is, with a P-value < 0.001), when the sample size is 100 and the true success rate is 70%.

Let's compare this to a ready-made function. For that we need to `install.packages("pwr")` as you may have done already at Learn R part of the course material.


```
#install.packages("pwr") #if you haven't done this yet, do it now
library(pwr) #load package "pwr"
#Compute power for binomial test at alpha = 0.001 and n = 100
#ES.h(0.5, 0.7) gives the "effect size" between null (0.5) and alternative (0.7) hypotheses
pwr.p.test(ES.h(0.5,0.7), sig.level = 0.001, n = 100)
```

```
##
##      proportion power calculation for binomial distribution (arcsine transformation)
##
##          h = 0.4115168
##          n = 100
##      sig.level = 0.001
##          power = 0.7952125
##      alternative = two.sided
```

This package gives power of ~80%, which agrees well with our manual calculation above. The `h` value in the output is the transformed effect size that corresponds to the difference between values 50% and 70%, and it was given by the `ES.h(0.5,0.7)` function call above.

We can leave also some other parameter than power away from the call to `pwr.p.test` and R will then compute the missing value for us. For example, we may want to know what n should be to get power of 90% at significance level 0.001.

```
#Compute n for given power by leaving out n from the parameters
pwr.p.test(ES.h(0.5,0.7), sig.level = 0.001, power = 0.9)
```

```
##
##      proportion power calculation for binomial distribution (arcsine transformation)
##
##          h = 0.4115168
##          n = 123.4389
##      sig.level = 0.001
##          power = 0.9
##      alternative = two.sided
```

The answer is that $n = 124$ would give power of 90% in this setting.

We can also ask values for several parameter values at a time. Let's compute power at the three significance levels 0.05, 0.001 and 1e-8.

```
#Compute power for three significance levels at once
pwr.p.test(ES.h(0.5,0.7), sig.level = c(0.05, 0.001, 1e-8), n = 100)
```

```
##
##      proportion power calculation for binomial distribution (arcsine transformation)
##
##          h = 0.4115168
##          n = 100
##      sig.level = 5e-02, 1e-03, 1e-08
##          power = 0.98442708, 0.79521249, 0.05309469
##      alternative = two.sided
```

We see that we have almost full power (98%) at level 0.05, but almost no power (5%) at level 1e-8.

Ingredients of statistical power

One-sample t-test power with `pwr.t.test()` Let's consider continuous data. We have a biomarker whose mean value in population studied (e.g. men aged 40-50 years) is 6.7 (sd = 1.0). We want to evaluate whether this biomarker is elevated in a particular subset of men (e.g. smokers). What is the power to detect 0.5 sd unit deviation from population mean at significance level 0.01 if we have collected 20, 50 or 100 individuals from the target population?

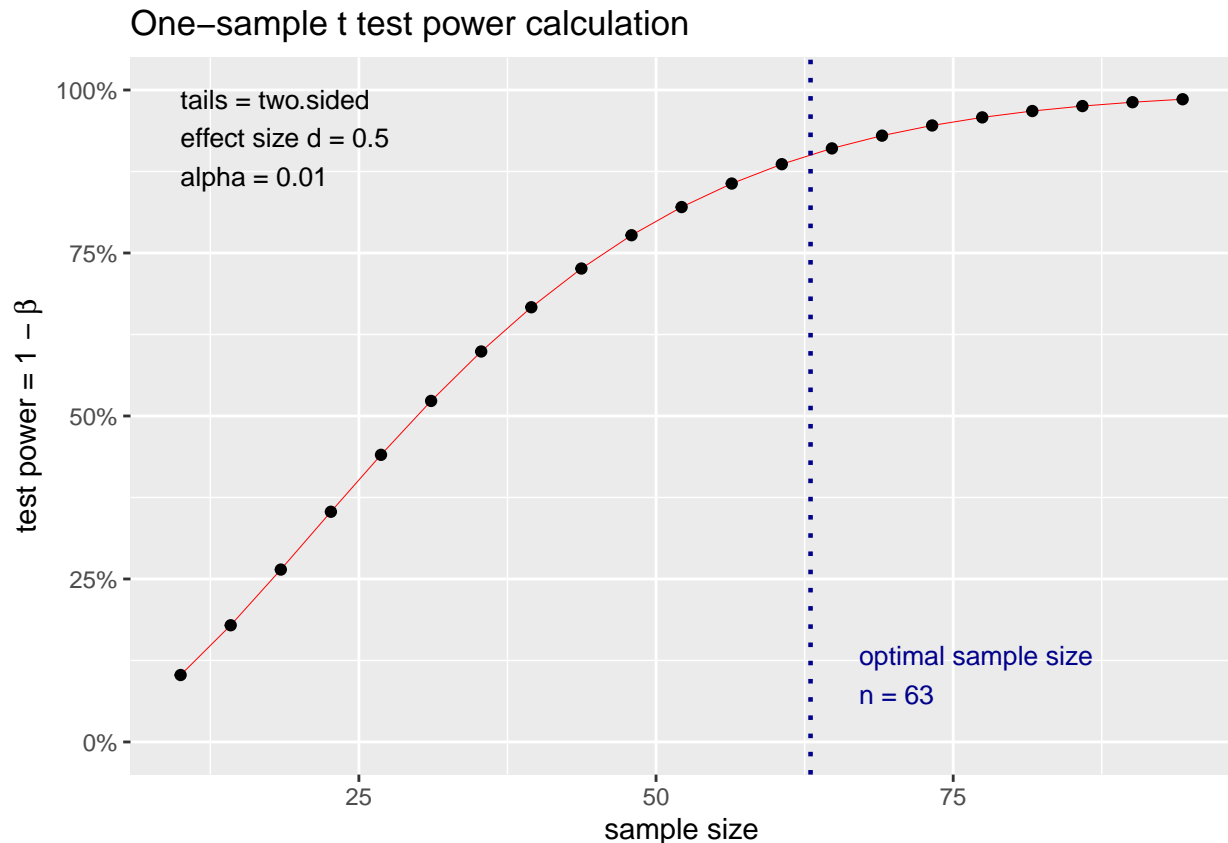
```
#We assume that sd in target group is the same as in the general population.  
# (Not necessarily true but how could we know otherwise before collecting data.)  
#Effect size for pwr.t.test is given as difference in means divided by  
# the standard deviation when it is the same in both groups  
# type="one.sample" means that we collect one sample and compare that to a KNOWN mean value.  
pwr.t.test(d = 0.5/1.0, sig.level = 0.01, n = c(20, 50, 100),  
           type = "one.sample", alternative = "two.sided")
```

```
##  
##      One-sample t test power calculation  
##  
##              n = 20, 50, 100  
##              d = 0.5  
##      sig.level = 0.01  
##      power = 0.2973461, 0.7993369, 0.9903473  
##      alternative = two.sided
```

So when both effect size (difference between the population mean and the subgroup mean) and the significance level (here 0.01) are kept fixed, we see that as n grows, it is easier to detect that the subgroup is different from the population mean with the statistical significance required.

We can use `pwr` package to plot a figure of sample size and power by applying `plot.power.htest()` to the output of `pwr` functions.

```
plot.power.htest(pwr.t.test(d = 0.5/1.0, sig.level = 0.01, power = 0.9,  
                           type = "one.sample", alternative="two.sided") )
```



Here “optimal sample size” corresponds to the value 90% given above in the function call.

If we keep n and significance level fixed and increase the effect (the difference in the means of the groups), then power will increase because larger true effects will generate data sets that look more and more extreme under the null, and hence will lead more easily to smaller P-values and more significant results than smaller true effects.

```
pwr.t.test(d = c(0.5, 1.0), sig.level = 0.01, n = 20,
           type = "one.sample", alternative = "two.sided")
```

```
##
##   One-sample t test power calculation
##
##           n = 20
##           d = 0.5, 1.0
##   sig.level = 0.01
##           power = 0.2973461, 0.9324954
##   alternative = two.sided
```

Finally, if we keep n and effect size fixed and ask for more and more stringent statistical significance, that is, results that would be more and more unlikely to be possible outcomes of chance effects under the null, then the power will decrease because we will accept only more and more extreme observations as being statistically significant.

```
pwr.t.test(d = 1.0, sig.level = c(0.01, 1e-6), n = 20,
           type = "one.sample", alternative = "two.sided")
```

```
##
##      One-sample t test power calculation
##
##          n = 20
##          d = 1
##      sig.level = 1e-02, 1e-06
##          power = 0.93249537, 0.04705335
##      alternative = two.sided
```

Examples of power calculations with pwr package

Let's go through some settings that pwr package handles. <https://cran.r-project.org/web/packages/pwr/vignettes/pwr-vignette.html>

Two sample proportion test with `pwr.2p.test()` We expect that a new treatment might reduce the proportion of recurrent infections from about 20% (which is the current rate) to about 10% in a specific hospitalized group of individuals. How large a randomised trial should we run in order to have good power to see this difference with significance level of 0.05? (Note that here we do not expect 20% to be exact value but will estimate the proportion also in the placebo group. This is different from earlier one sample proportion test that compared to the fixed null value of 50% using `pwr.p.test()`.)

```
pwr.2p.test(h = ES.h(0.2, 0.1), power = 0.90, sig.level = 0.05)
```

```
##
##      Difference of proportion power calculation for binomial distribution (arcsine transformation)
##
##          h = 0.2837941
##          n = 260.9272
##      sig.level = 0.05
##          power = 0.9
##      alternative = two.sided
##
## NOTE: same sample sizes
```

So we would need about 261 individuals in both the treatment and in the placebo group.

What if we had a fixed size of the treatment group of $n_1 = 230$ set by available resources, but could collect a larger placebo group. How large would the placebo group need to be for 90% power at significance level of 0.05?

```
pwr.2p2n.test(h = ES.h(0.2, 0.1), n1 = 230, power = 0.90, sig.level = 0.05)
```

```
##
##      difference of proportion power calculation for binomial distribution (arcsine transformation)
##
##          h = 0.2837941
##          n1 = 230
##          n2 = 301.4638
##      sig.level = 0.05
##          power = 0.9
##      alternative = two.sided
##
## NOTE: different sample sizes
```

We would need about $n_2 = 302$ individuals in the placebo group.

Chi-square test for contingency table with `pwr.chisq.test()` Suppose that patients are treated with certain therapy and the outcome is recorded as Good, Satisfactory or Bad. It is observed that about 20% of people fall into Bad category and about 40% to Good category. MD expects that the treatment works better for men than for women and assumes that the proportion of men in Bad is 30% while for women it is only 10%. However, the Good category seems similar in both sexes. How large a sample (with equal number of men and women) would one need to see this effect at significance level 0.01 using a Chi-square test for a 2x3 table?

```
#Let's generate 2x3 matrix of probabilities for the alternative hypothesis.
```

```
p.alt = matrix(c(0.5*0.4,0.5*(1-0.1-0.4),0.5*0.1, #1st row is for women
                0.5*0.4,0.5*(1-0.3-0.4),0.5*0.3),
              dimnames=list(c("M","F"),c("Good","Satisf","Bad")),
              byrow=TRUE, ncol=3, nrow=2)
p.alt
```

```
##   Good Satisf  Bad
## M  0.2   0.25 0.05
## F  0.2   0.15 0.15
```

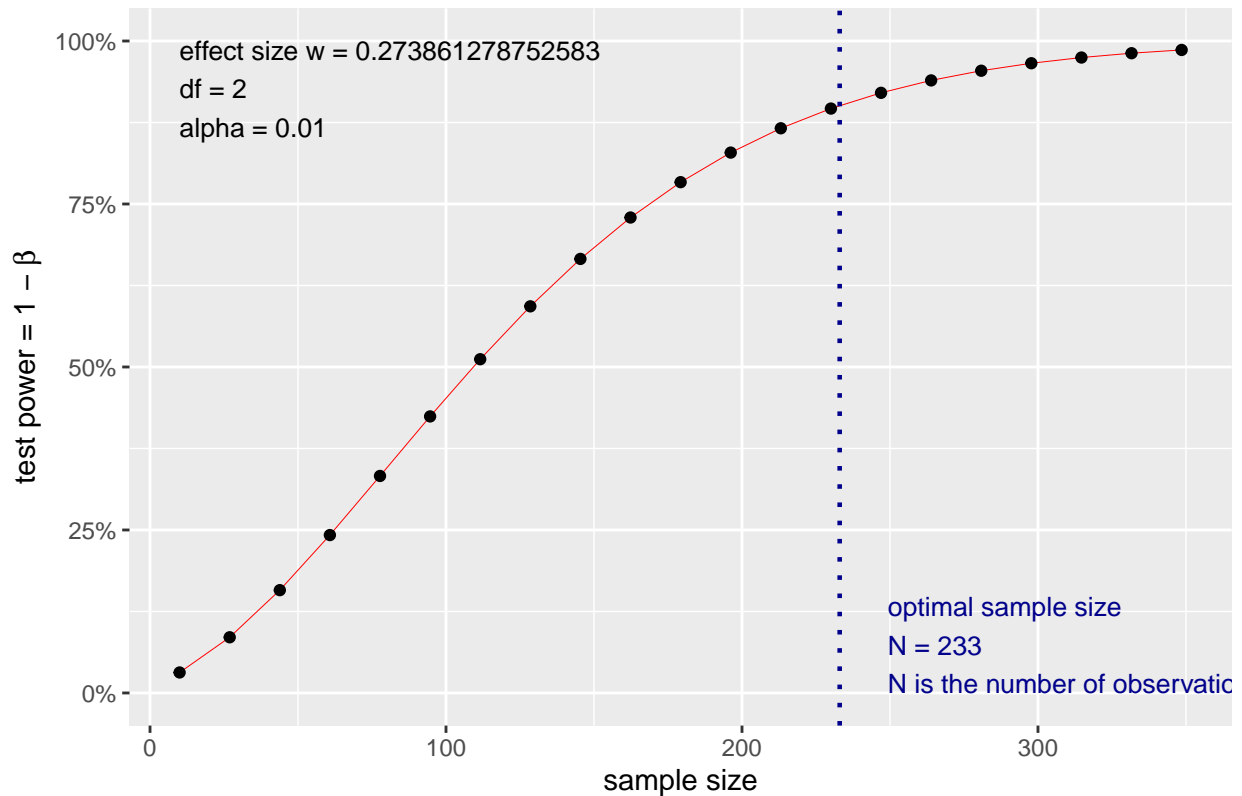
```
#effect size for chi-square test is given by ES.w2 function applied to p.alt
#degrees of freedom of chi-square test if (Rows-1)x(Cols-1), here (2-1)x(3-1)
pwr.chisq.test(ES.w2(p.alt), df=(2-1)*(3-1), sig.level = 0.01, power=0.90)
```

```
##
##   Chi squared power calculation
##
##           w = 0.2738613
##           N = 232.3559
##           df = 2
##   sig.level = 0.01
##           power = 0.9
##
## NOTE: N is the number of observations
```

```
#We would need about 240 individuals, 120 males and 120 females.
```

```
plot.power.htest(pwr.chisq.test(ES.w2(p.alt), df=(3-1)*(2-1), sig.level = 0.01, power=0.90))
```

Chi squared power calculation



One sample t-test with `pwr.t.test(, type = "one.sample")` We want to test whether a group treated with medication M shows a decrease in risk factor R. Suppose that the mean in the risk population is 4.1 (sd 0.8) and we measure a group treated with M 12 months after starting the therapy. How large a group needs to be in order to have 90% power to find a difference in mean of 0.3 at 2-sided significance level of 0.05?

```
#NOTE: use "one.sample" to compare a group to a fixed reference value (here pop mean of 4.1)
# and give effect size parameter d in units of sd of population (here 0.8)
pwr.t.test(d = 0.3 / 0.8, sig.level = 0.05, alternative = "two.sided",
           power = 0.9, type = "one.sample")
```

```
##
## One-sample t test power calculation
##
## n = 76.66599
## d = 0.375
## sig.level = 0.05
## power = 0.9
## alternative = two.sided
```

Two sample t-test with equal sample sizes with `pwr.t.test(, type = "two.sample")` We do the same comparison but collect equal sized groups from the risk population and from the treatment group. How large groups?

```
#NOTE: use "two.sample" to compare two groups
pwr.t.test(d = 0.3 / 0.8, sig.level = 0.05, alternative = "two.sided",
           power = 0.9, type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##              n = 150.4057
##              d = 0.375
##      sig.level = 0.05
##              power = 0.9
##      alternative = two.sided
##
## NOTE: n is number in each group
```

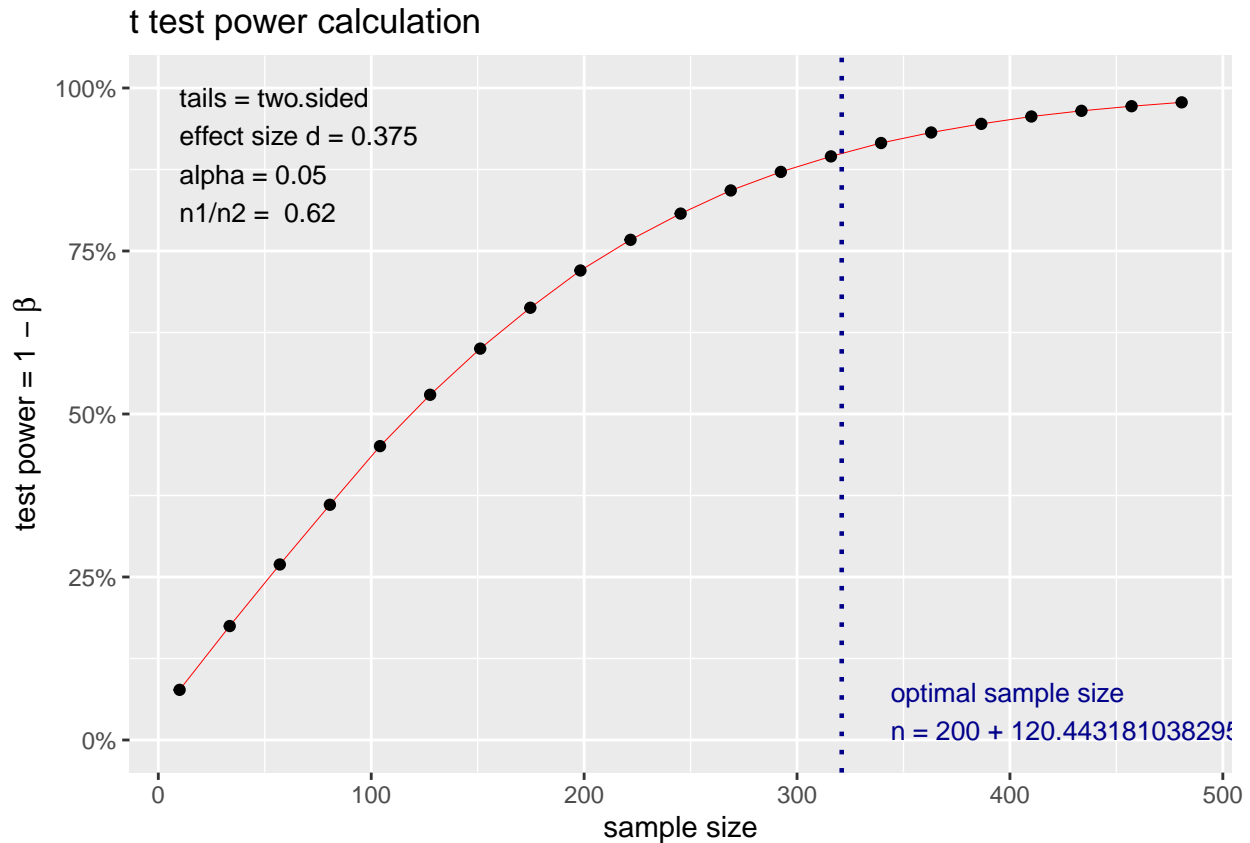
Note how we need double the sample size in each group compared to the one-sample test above. This is because now we need to estimate means of two groups and they both come with some uncertainty that further accumulates in the calculation of the difference between the groups. Hence we need larger samples here than when comparing an estimated mean of one group against a fixed value.

Two sample t-test with unequal sample sizes with `pwr.t2n.test()` What if we had a fixed set of 200 individuals from the risk group measured and need to collect a sample from the treatment group. How large a group?

```
#NOTE: We use "pwr.t2n.test" to have different n in the two groups
pwr.t2n.test(d = 0.3 / 0.8, sig.level = 0.05, alternative = "two.sided", power = 0.9, n1 = 200)
```

```
##
##      t test power calculation
##
##              n1 = 200
##              n2 = 120.4432
##              d = 0.375
##      sig.level = 0.05
##              power = 0.9
##      alternative = two.sided
```

```
plot.power.htest(
  pwr.t2n.test(d = 0.3 / 0.8, sig.level = 0.05, alternative = "two.sided", power = 0.9, n1 = 200))
```



Example 5.3. *Selective laser trabeculoplasty versus eye drops for first-line treatment of ocular hypertension and glaucoma (LiGHT): a multicentre randomised controlled trial* in Lancet 2019 compares two treatments (laser trabeculoplasty and eye drops) for patients with open angle glaucoma or ocular hypertension and no ocular comorbidities. The primary outcome was health-related quality of life at 3 years, assessed by EQ-5D. They write: “We calculated that a sample size of 718 patients was needed to detect a difference of 0.05 in EQ-5D between the two groups using a two sample t test at the 5% significance level with 90% power, assuming a common standard deviation of 0.19 and 15% attrition.”

Let’s see if we agree. Here the group difference should be input in units of SD and is $d=0.05/0.19$.

```
pwr.t.test(d = 0.05/0.19, power = 0.9, type = "two.sample",
           alternative = "two.sided", sig.level = 0.05)
```

```
##
##      Two-sample t test power calculation
##
##              n = 304.4178
##              d = 0.2631579
##      sig.level = 0.05
##      power     = 0.9
##      alternative = two.sided
##
## NOTE: n is number in *each* group
```

So each group should have 305 individuals and hence in total we need $2 \cdot 305 = 610$ individuals. If we take into account that 15% of the recruited individuals may leave the study (15% attrition rate), we need to recruit a larger group n for which $0.85n = 610$. Thus we need


```
ceiling(610/0.85) #ceiling function rounds up the values to next integer
```

```
## [1] 718
```

individuals, just like they reported.

Power in designing an experiment

If the experiment is not powerful to answer the question of interest, then we are unlikely to learn much from it. And even worse, if do not understand that a non statistically significant effect may be because of small power, we may wrongly conclude that it has been “statistically” shown that the effect does not exist (see below). We should avoid designing underpowered studies.

Power when interpreting the results

In a cohort of 200 female appendicitis cases a new scoring system was found to predict the severity of the condition: difference between averages of severe and benign groups were 10 (5.2, . . . ,14.0) with P-value of 0.001. You collected 30 male cases and saw that the corresponding values were 7 (-5.0, . . . ,20.4) P=0.56. Does this mean that the appendicitis scoring system works well for women but not for men because in women the difference was statistically significant at level 0.001 but in men it was not significant even at level 0.05?

If our experiment is underpowered to answer the question of interest, then a negative result (the one that does not find a statistically significant difference between groups/treatments) does NOT mean that there is no difference. “Absence of evidence is not evidence of absence.” <http://www.bmj.com/content/311/7003/485.full>

If we have designed an experiment with good power to detect an effect, but still don’t see the effect, then we can conclude that the effect (of that size) does probably not exist. However, smaller effects may still well exist.

If we have really large studies, we can find statistically significant effects between almost any two groups. The more relevant question is that how large effects have any practical importance. For example, even if we can show by using very large samples that the prevalence of MS-disease in people born in April (say risk of 11/10000) is statistically significantly higher than those born in October (say risk of 10/10000) this does not have much practical importance since the increase in absolute risk is tiny. Scientifically though, such observations may be interesting and open up new avenues for understanding the disease.