Lecture 2: P-value

Matti Pirinen

8.9.2021

In the previous lecture, we learned to use the cumulative distribution function `pbinom()` of the binomial distribution to compute tail probabilities of the binomial distribution. Such tail probabilities answer to questions such as what is the probability that (i) at most 10 patients survive or (ii) at least 20 patients survive, out of 50 patients when the average survival rate is 30%. In this lecture, we use such tail probabilities to define the concept of P-value that is typically presented with every statistical test in medical literature.

Definition of P-value

Suppose we have 100 pairs of same-sex twins that are discordant for psoriasis (i.e. one has psoriasis and the other doesn't) and we know that BMI is higher for the psoriatic individual rather than his/her non-psoriatic twin sibling in 71 out of 100 pairs. Does that sound like higher BMI has a tendency to go with the psoriasis status in these twin pairs? How to statistically quantify association between BMI and psoriasis in these data?

Idea: If psoriasis was independent of BMI, then we would expect that there is a 50% chance that the individual with psoriasis happens to have a higher BMI than his/her twin sibling. In other words, the number of twin pairs where the psoriatic one has higher BMI is distributed as Bin(100, 0.5), i.e, has binomial distribution with 100 trials each having a success probability of 0.5. Think this as the distribution of tails (suom. klaava) in 100 tosses of a fair coin where each toss corresponds to a twin pair. Under the independence assumption, we would expect about 50 successes (out of 100 trials) in this binomial experiment, but with some variation around the exact value 50. And we did observe 71. Is 71 already so far from 50 that we have a reason to suspect the independence assumption? Or could it easily happen "by chance" if success probability is 0.5? How can we tell?

We want to quantify (or "test") how consistent our observation is with the independence assumption. In statistics terms, we define a **null hypothesis** (denoted by H_0) saying that "psoriasis status is independent of BMI". This can be expressed as

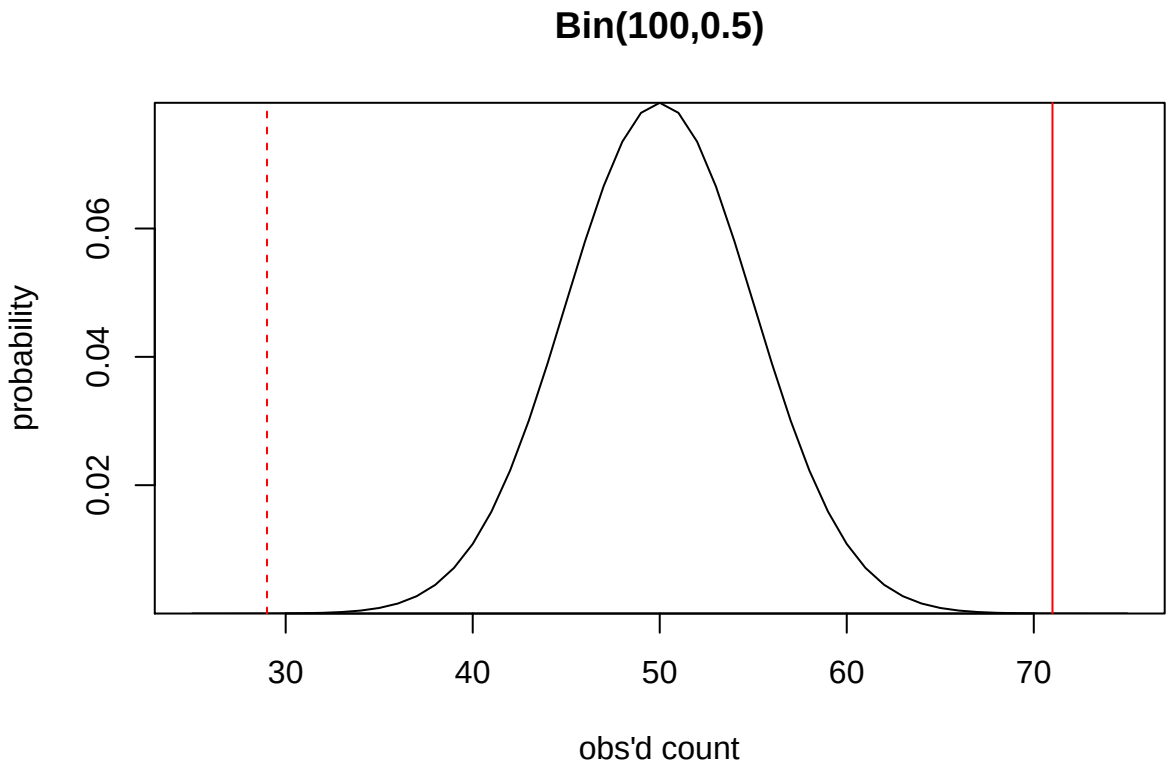
$$H_0 : p = 0.5,$$

which says that the null hypothesis (H_0) is that the probability p that the psoriatic twin has a higher BMI than his/her twin sibling is 50%. Often the null hypothesis is defined in such a way that it describes the default situation which, if it holds true, is not very interesting or surprising. Instead, if the null hypothesis turns out not to hold true, that is usually interesting or surprising. A standard way to test the null hypothesis is by quantifying how likely the observed data are under the null hypothesis. The logic is that if the kind of data that we have observed are very unlikely under the null hypothesis, then we have good reasons to think that the null hypothesis might not be true.

Let's plot the probability distribution of data under the theoretical null hypothesis (that assumes psoriasis and BMI are independent). We could use barplot like in Lecture 1 but now we will rather make a continuous curve that joins the probabilities together using the `plot()` function.

```r
n = 100 #pairs of twins
#Let's cut to range 25...75 as tails outside this are very improbable and hence unnecessary to show her
x = 25:75
y = dbinom(x, size = n, prob = 0.5)
#Let's plot points in x-y coordinates where x is number of twin pairs 25,...,75
# for which psoriasis and higher BMI occur in same individual
# and y is probability of that observation under the null hypothesis.
plot(x, y, xlab="obs'd count", ylab = "probability", main = "Bin(100,0.5)", t = "l", yaxs = "i")
#main = gives the title for the plot. it should be in quotes " "
#t="l", means that plotting type is "line", that is, points are connected by line.
#yaxs="i" forces plotting area strictly to the range of observed values on y-axis

#Let's also mark our real observation that was 71.
x.obs = 71
#let's add a vertical line to our observation using abline()
abline(v = x.obs, col = "red") #v = x.obs means vertical line at x.obs
abline(v = n - x.obs, col = "red", lty = 2) #lty = 2 makes a dashed line at 29 = 100 - 71
```



The observation 71 is very much at the right hand tail of the null distribution, so it is pretty improbable to get this many or even more observations under the null hypothesis (that is, if BMI and psoriasis were independent).

Compute probability that under the null hypothesis one would get an observation at least as far in the right tail than we have observed.

```r
#pbinom(70, n, 0.5) reports probability mass cumulated from 0 to 70,
#and '1-pbinom(70, n, 0.5)' is then the mass cumulating from 71 to 100.
1 - pbinom(x.obs - 1, size = n, prob = 0.5)
```

```
## [1] 1.608001e-05
```

So in less than 2 out of 100,000 experiments under the null hypothesis would we get 71 or more observations. What's your conclusion about independence between psoriasis and BMI?

Conclusion: Either psoriasis status and higher BMI are linked, or a very improbable event has happened "by chance", namely that we have observed at least as many as 71/100 twin pairs where high BMI and psoriasis go together, which is already so far from 50%:50% expectation that it has a very small probability if BMI and psoriasis were independent.

Let's list these right-hand tail probabilities for even more extreme possible outcomes

```
#Let's show values 71,...,100 and combine them with corresponding tail probabilities
# using cbind() function, "column bind"", combines columns into a matrix.
x = x.obs:n
cbind(x, 1 - pbinom(x - 1, size = n, prob = 0.5) )
```

```
##          x
##  [1,]  71 1.608001e-05
##  [2,]  72 6.289575e-06
##  [3,]  73 2.346206e-06
##  [4,]  74 8.336813e-07
##  [5,]  75 2.818141e-07
##  [6,]  76 9.050013e-08
##  [7,]  77 2.756790e-08
##  [8,]  78 7.952664e-09
##  [9,]  79 2.168683e-09
## [10,]  80 5.579545e-10
## [11,]  81 1.351381e-10
## [12,]  82 3.073897e-11
## [13,]  83 6.548984e-12
## [14,]  84 1.302958e-12
## [15,]  85 2.412515e-13
## [16,]  86 4.141132e-14
## [17,]  87 6.550316e-15
## [18,]  88 8.881784e-16
## [19,]  89 1.110223e-16
## [20,]  90 0.000000e+00
## [21,]  91 0.000000e+00
## [22,]  92 0.000000e+00
## [23,]  93 0.000000e+00
## [24,]  94 0.000000e+00
## [25,]  95 0.000000e+00
## [26,]  96 0.000000e+00
## [27,]  97 0.000000e+00
## [28,]  98 0.000000e+00
## [29,]  99 0.000000e+00
## [30,] 100 0.000000e+00
```

We see that probabilities get smaller and smaller as we would observe more extreme deviations from value 50 (that is the expectation under the null hypothesis). Intuitively, more extreme observations also give more evidence that the deviation from 50 is not just a "chance event" or a "statistical fluctuation" but rather there is some real link between BMI and psoriasis. This is because probability of getting so extreme values purely by "chance" goes essentially to 0.

P-value We have just encountered a "P-value". Definition of **P-value is the probability of getting "at least as extreme dataset under the null hypothesis as was observed"**. The logic is that if

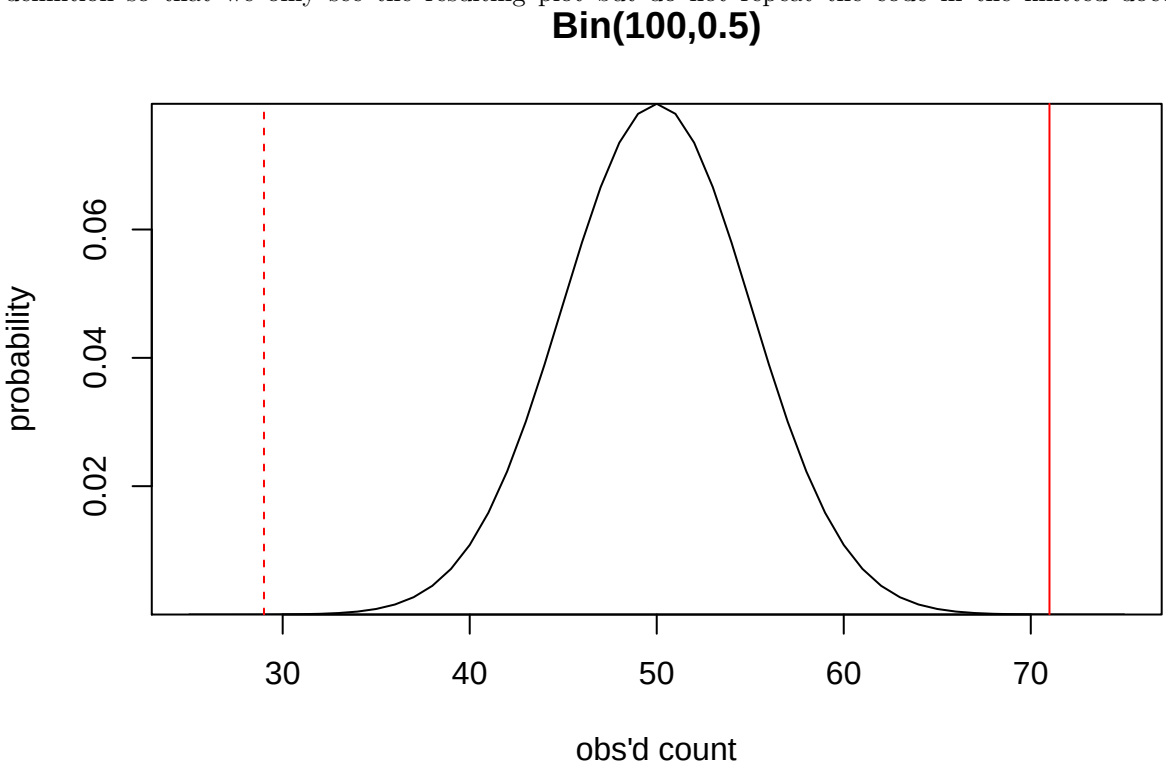
P-value is very small, then it would be very improbable to observe a data set that is at least as extreme as was observed if the null hypothesis was true. Therefore an observation with small P-value may make us to question the null hypothesis.

Two questions emerge:

(1) How do we define which data sets are "at least as extreme" as the observed one under the null hypothesis?

(2) How small a P-value should be in order that we can conclude that the null hypothesis is not true? Or can we ever do that?

Which data sets are "at least as extreme" as the observed one? Let's repeat the previous density plot of the possible outcomes in range 25 to 75 of a binomial experiment Bin(100, 0.5) with the two red lines at 71 and 29. (Let's suppress the code from output by using `echo = FALSE` in the code block definition so that we only see the resulting plot but do not repeat the code in the knitted document.)



In this example, the most probable value of twin pairs where higher BMI and psoriasis occur in the same individual is 50 (out of 100) under the null hypothesis that BMI and psoriasis are independent. It seems intuitive that the larger the deviation from 50, the more "extreme" the observation is under the null hypothesis. A quantitative definition could be that out of two putative observations, the one with the smaller probability (under the null) is the "more extreme" one.

Above we computed that probability under the null to observe 71 or more pairs is about 2 in 100,000. If we had not fixed beforehand that we would only be interested in deviations that are towards higher values from 50, then observing at least 71 out of 100 would be as extreme value from 50 as observing at most 29 out of 100 (dashed red line above). So probability under the null of seeing "as extreme data" as has been observed would mean observing "at least 71 or at most 29 cases out of 100 trials". Let's compute the P-value under such a **two-sided alternative hypothesis**.

```r
pbinom(29, size = n, prob = 0.5) + (1 - pbinom(71-1, size = n, prob = 0.5))
```

```
## [1] 3.216002e-05
```

This is a two-sided P-value (no prior hypothesis which direction defines more extreme cases so we consider both directions) and gives here 2 x P-value from the one-sided test that would consider only one tail of the distribution. Typically, we consider 2-sided tests since we want to see whether there is any deviation from the null hypothesis no matter to which direction. One-sided test ignores the other direction and is conceptually applicable only rarely when we have a strong reason to pre-define the tail of interest, e.g., if we try to replicate a result from an earlier study.

More about one-sided vs two-sided tests in BMJ 1994;309:248 http://www.bmj.com/content/309/6949/248.full.

How small a P-value should be in order that we can conclude that the null hypothesis is not true? The short answer is that we cannot ever conclude truth based on statistics alone. And that even to properly quantify the level of evidence that the null hypothesis is true / not true, we would need other information in addition to the P-value.

However, numerical value of P-value does tell something and, very generally, smaller P-values are more evidence against the null hypothesis than larger P-values. This is linked to the technical term of **statistical significance**, which is used when a statistical test has produced a certain pre-defined level of evidence against the null hypothesis. Often this evidence is measured by the P-value where smaller P-values correspond to "more statistically significant" results.

For example, based on the previous psoriasis-BMI example, we could conclude that there is a statistically significant association between psoriasis and BMI, at a **significance level** of 0.001, because the P-value was about 3e-5 $<$ 0.001. The association is also significant at the level of 1e-4 (because 3e-5 $<$ 1e-4), but not anymore at the level of 1e-5 (because 3e-5 $>$ 1e-5). The most common significance level in medicine, and elsewhere in applied statistics, is 0.05, which is unfortunate, since a P-value around 0.05 is typically not at all strong evidence against the null hypothesis. This is a problem because it means that many "statistically significant" results from literature are actually false positives, where P-value simply has got below 0.05 "by chance" even though the null hypothesis holds. Note that the probability of having a P-value smaller than 0.05 "by chance" even when the null hypothesis holds is 5% or 1 in 20, so it indeed happens quite often just by chance even when the null hypothesis is true.

All thresholds for calling P-values significant (such as the commonly used 0.05) are arbitrary and they should be only one piece of evidence when we determine which hypothesis is the most likely to hold. A bad practice is to dichotomize P-values by only reporting whether they were "significant/non-significant" at a certain significance level rather than reporting the exact numerical value of the P-value. Unfortunately, the pursue of "statistical significance" is often so strong in medical literature that the other, often more important, aspects of the statistical analyses are not paid enough attention. Partly this is because in order to get publications, one needs to find some "statistically significant" results. Consequently, there are a lot of published studies where P-value has just got slightly below 0.05, but when a similar experiment is repeated again, no supporting evidence for the claimed result can be found, possibly because the results was a false positive in the first place.

Another important aspect in medical field is that even if there is a true statistically significant effect, it may not be significant for clinical practice: https://www.students4bestevidence.net/statistical-significance-vs-clinical-significance/.

Summary: P-value is the probability that if the null hypothesis was true, then we would get at least as extreme data set as what we have observed. Its purpose is to quantify the possible inconsistency between the null hypothesis and the observed data. If P-value is small, then it seems unlikely that under the null hypothesis this extreme data would had occured. Thus small P-values give us a reason to suspect that null

hypothesis may not be true. However, it may still be the case that the null hypothesis is true but a rare event has occured and hence we see a small P-value.
Statistical significance at level α is a technical term that is used when the P-value falls below a pre-defined significance level α, but it does not conclusively show that the null hypothesis does not hold, and P-values should be reported as they are, not dichotomize based on significance level.

IMPORTANT: P-value is always a probability under the null hypothesis of observing certain kind of data sets, it is NOT a probability about the null hypothesis itself. Thus, based on the P-value alone, you cannot say what is the probability that the null hypothesis holds. All you can say with the P-value is that if the null hypothesis holds, then the probability of getting data like these or even more extreme, is given by the P-value.

Want to read another explanation of P-value from BMJ? https://www.bmj.com/content/340/bmj.c2203.

Example 2.1

(1) You want to study whether the appendicitis (umpilisäke) cases operated at Meilahti Hospital have any bias from 50%-50% split between men and women. What is your null hypothesis?

Null hypothesis is that there is no deviation from the 50%:50% setting, in other words, that males and females are equally susceptible to appendicitis. (An underlying assumption here is that there are equal number of males and females in the population, which is not true, for example, for all age groups, but we are happy with this assumption for now.) If we would think this setting as a binomial experiment, then given that we observe n patients, we would assume that the number of male patients among the observed patients would be distributed as Bin(n, 0.5) if the null hypothesis holds.

(2) Over the last year, there has been 316 patients of which 180 women and 136 men. What is the P-value under the null hypothesis? Is the proportion of men statistically significantly different from 50% at significance level 0.05? Do you suspect a bias from 50%-50% proportions based on this P-value?

```
n = 316 #all patients
x = 136 #men
x/n #proportion men is somewhat < 0.5
```

```
## [1] 0.4303797
```

```
#Let's compute 2-sided P-value. Since x/n < 0.5, we compute left tail and multiply by 2
2*pbinom(x, size = n, prob = 0.5) #2-sided P-value
```

```
## [1] 0.0154331
```

P-value is statistically significant at level 0.05 (since 0.015 $<$ 0.05). This result could indicate a bias where the population of patients has less men than women, however, the statistical evidence is not yet very strong (a P-value of 0.015 happens even under the null more often than once in a hundred experiments by chance), and more data would be useful to get more convincing evidence to either direction.

(3) If you do a comparison between treatment and placebo group and statistics show a difference with a P-value of 0.003, what does that mean exactly, explained in plain language? What about if the P-value were 0.00008? Which one of these two P-values would show stronger evidence for a true treatment effect and why?

P-value of 0.003 from a clinical trial means that we have observed a difference of such a magnitude that it or a more extreme difference would be observed only in 3 out of 1000 clinical trials like this, if the null hypothesis holds (i.e. if there is no true difference between treatment and placebo). So we tend to think that it is quite improbable (0.003) to observe at least this extreme data "by chance" when the null hypothesis holds. (But still, it can happen, even under the null hypothesis.)

IMPORTANT: It would be WRONG to say that probability of this observation being produced "by chance" is 0.003. That would suggest that 0.003 was the probability of the null hypothesis while it actually is a probability of certain sets of data under the null hypothesis, and it is NOT a probability of the null hypothesis. Based on only the P-value, it is never possible to say how probable a null hypothesis is.

P-value of 0.00008 is clearly < 0.003, and it does provide more evidence than P-value of 0.003 that there is a true treatment effect. This is because typically a more extreme deviation from the expectation under the null hypothesis is more evidence against the null hypothesis, even though we cannot derive the exact probability for the null hypothesis based on the P-value alone.

Binomial test and Fisher's test

`binom.test()` Above we have used the binomial distribution function `pbinom()` to compute P-values. In practice, we can do the same by using a ready-made function `binom.test()`.

Let's redo the statistical testing of 71 successes out of 100 trials when the null hypothesis is that success probability is 0.5. (We hope to get 3.216e-5 to be consistent with our earlier result.)

```
binom.test(71, n = 100, p = 0.5)
```

```
##
##  Exact binomial test
##
## data:  71 and 100
## number of successes = 71, number of trials = 100, p-value = 3.216e-05
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.6107340 0.7964258
## sample estimates:
## probability of success
##                   0.71
```

Output includes the input data (successes and trials) and the P-value under the null hypothesis of $p = 0.5$, when P-value is computed using the two-sided test ("**alternative hypothesis**" is that value can be different from 0.5 to **either** direction).

We could also do a one-sided test, where the alternative hypothesis is that the observed value is greater than the expected under the null. To do this, we use `alternative = "greater"` in `binom.test()`.

```
binom.test(71, n = 100, p = 0.5, alternative = "greater")
```

```
##
##  Exact binomial test
##
## data:  71 and 100
## number of successes = 71, number of trials = 100, p-value = 1.608e-05
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
```

```
##  0.6262716 1.0000000
## sample estimates:
## probability of success
##                   0.71
```

Note how the "alternative hypothesis" has changed in the output and now only the upper tail probability is computed. Alternatively, only the lower tail would be considered if `alternative = "less"`. If `alternative=` is not specified, it is taken to be `two.sided` by default.

Type `?binom.test` to see the details of how to use this function.

Example 2.2

(1) Over the last year, there has been 316 appendicitis patients of which 180 women and 136 men. Make a binomial test to see whether this is a significant difference from the null hypothesis that there are equal number of male and female patients at significance level 0.05.

```
binom.test(136, n = 316, p = 0.5, alt = "two.sided") #136 successes out of 316 trials when prob = 0.5
```

```
##
##  Exact binomial test
##
## data:  136 and 316
## number of successes = 136, number of trials = 316, p-value = 0.01543
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.3750857 0.4869912
## sample estimates:
## probability of success
##              0.4303797
```

As seen in example 2.1, this is a statistically significant results at significance level 0.05.

(2) What if the female population in the area was larger (60% of people) than the male population (40% of people). Now we would expect more females among patients even under the null hypothesis of similar appedicitis incidence between sexes. Namely, now the null hypothesis of equal incidence in males and females says that, for a patient, the probability of being male is 0.4.

```
binom.test(136, n = 316, p = 0.4, alt = "two.sided")
```

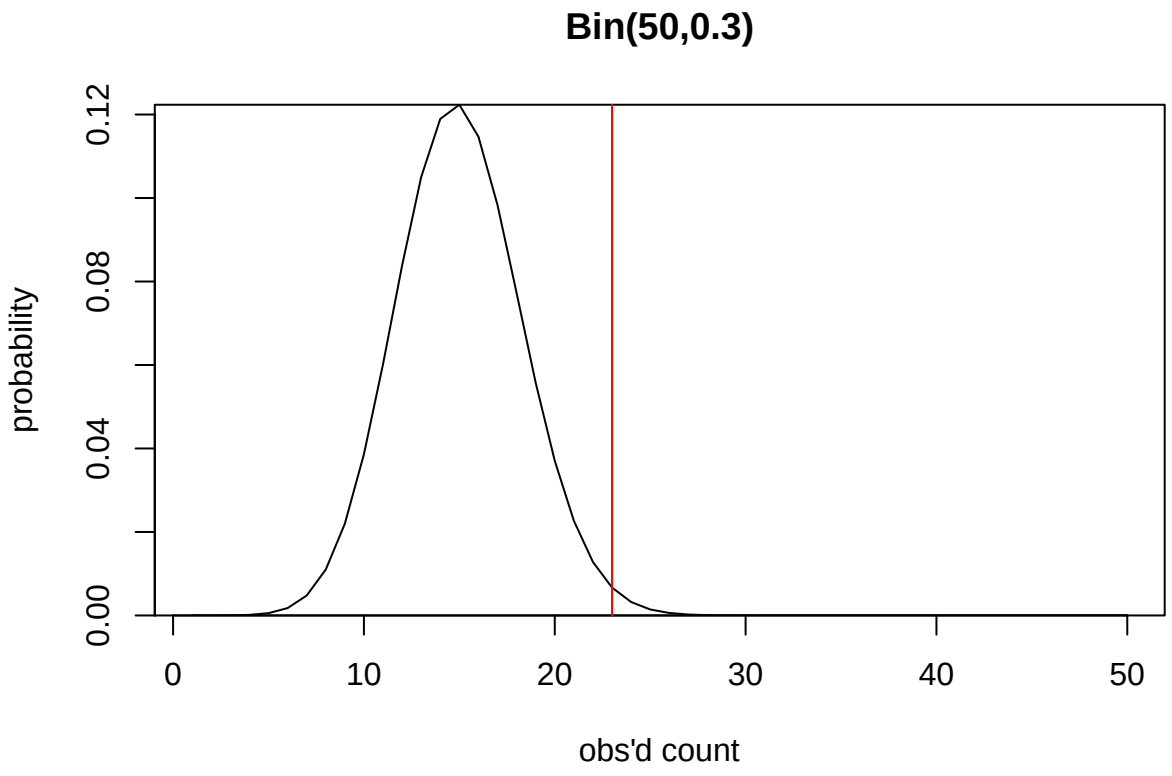
```
##
##  Exact binomial test
##
## data:  136 and 316
## number of successes = 136, number of trials = 316, p-value = 0.2756
## alternative hypothesis: true probability of success is not equal to 0.4
## 95 percent confidence interval:
##  0.3750857 0.4869912
## sample estimates:
## probability of success
##              0.4303797
```

We see that P-value is larger now (> 0.25) and we would have no reason to doubt the null hypothesis of equal incidence in males and females based on these results. Thus, we conclude that an unequal (60%:40%) population size between sexes would already be enough to explain the apparent difference of 180:136 in appendicitis cases, since there is not anymore any statistically significant deviation from the null hypothesis once the null hypothesis takes into account the ratio between males and females.

Example 2.3 Suppose that you are treating 50 patients and you expect that the success rate is 30%. You observe 23 successes and 27 failures. What is the two-sided P-value for the observation under the null hypothesis that the success rate is 30%? Do these data make you question the null hypothesis?

Answer: We expect a rate of 30% and we have observed an empirical rate of $0.46 = 23/50$. Let's draw a picture of the null distribution and the observation, adapting the code we used earlier for distribution Bin(100,0.5).

```
n = 50
p.null = 0.3
x = 1:n
y = dbinom(x, size = n, prob = p.null)
plot(x, y, xlab="obs'd count", ylab = "probability", main = "Bin(50,0.3)", t = "l", yaxs = "i")
x.obs = 23
abline(v = x.obs, col = "red")
```



We see that probability of getting at least this large counts under the null is the right-tail probability:

```
p.right = 1 - pbinom(x.obs-1, size = n, prob = p.null) #right-hand sided P-value
p.right
```

```
## [1] 0.01227613
```

What would be the 2-sided P-value? As opposed to our earlier example with Bin(100, 0.5), this distribution, Bin(50, 0.3), is not symmetric with respect to its expected value 15, and thus there is not (necessarily) any value < 15 such that the left-tail probability at that value would be exactly `p.right`=0.123. Thus we need to choose among two options to determine a two-sided P-value:

(1) Sum up the probabilities of all those outcome values whose probabilities are less than or equal to the probability of the observed value. This is an exact way to define a two-sided P-value, and this is how `binom.test()` function does it.

```
binom.test(x.obs, n = n, p = p.null)
```

```
##
##  Exact binomial test
##
## data:  x.obs and n
## number of successes = 23, number of trials = 50, p-value = 0.01954
## alternative hypothesis: true probability of success is not equal to 0.3
## 95 percent confidence interval:
##  0.3181492 0.6067580
## sample estimates:
## probability of success
##                   0.46
```

(2) Simply multiply the right-tail probability by two to represent the idea that we want to reserve an equal amount of probability from the left tail for "at least as extreme observations" as we do from the right tail. This simple approximation works fine for practical purposes, but does not give exactly the same P-value as given by `binom.test()`.

```
2*p.right
```

```
## [1] 0.02455225
```

Both methods give a P-value of about 0.02. Conclusion is that there is some evidence that the success rate might be larger than 30% (statistically significant deviation from the null at significance level 0.05), but more data would still be needed before evidence for the deviation might become very convincing.

`fisher.test()` In binomial test we had a fixed null hypothesis parameter value (like $p = 0.5$) and then we measured one property on each individual (e.g. being male or female). Going one step further, we may have our observations characterized by two properties, for example, being male or female and having the disease or being healthy. How can we test whether these two properties are independent of each other or whether they are associated with each other. Here independence between sex and disease means that the disease incidence is the same in both sexes, i.e., it does not depend on the sex. And an association between sex and disease means that disease is more common among one of the sexes than among the other sex. (In large samples, we can compute the disease incidence in each sex and compare those with proportion test, as we will do in next lecture. For small samples, we should use Fisher's test, as we do now.)

As an example, suppose we measure the carrier status of a particular rare genetic variant of the HLA-C gene on the immune system related HLA region of the human genome in healthy controls and in psoriasis patients, and we get the following counts.

	non-carriers	carriers
healthy	40	1
psoriasis	13	6

The question is whether the variant carrier status is associated with the disease status, or whether the two properties are independent of each other.

We can test this by Fisher's test. In R, that works by making a 2x2 -matrix or data frame of data and applying `fisher.test()` on it. Let's use data frame and add column and row names.

```
x = data.frame(noncar = c(40,13),
               carrier = c(1,6),
               row.names = c("healthy", "psoriasis"))
x #always check that your data matrix looks correct before analysis
```

```
##           noncar carrier
## healthy       40       1
## psoriasis     13       6
```

```
fisher.test(x)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  x
## p-value = 0.003011
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    1.864826 865.022378
## sample estimates:
## odds ratio
##    17.4474
```

Fisher's exact test returns a P-value as the sum of probabilities of all such 2x2 tables that have the same row and column counts as we have observed, and that have at least as small probability as the one we have observed. Thus here the "at least as extreme" criteria of the P-value definition is defined by how probable each 2x2 table is if we were to keep the row and column sums fixed but otherwise were to randomly assign the numerical values to the 4 cells of the table. You can read more from Wikipedia.

Our observation is that only 1/41 healthy people have the variant but 6/19 patients have it. The P-value describes how statistically extreme this observation is. The P-value 0.003 says that only in 3 out of 1000 random tables with these total counts, the genetic variants are distributed at least in this extreme way between the two groups.

Example 2.4 Of 30 people employed in a small workshop 18 worked in one department and 12 in another department. In one year five of the 18 reported hand injury, and of the 12 men in the other department one did so. Is there a difference in the departments and how would you report this result?

In essence, we are comparing whether proportions 5/18 and 1/12 are statistically different, and we do that by Fisher's test. Now we use matrix rather than data frame.

```r
x = rbind(c(18-5,12-1),c(5,1)) #1st row: healthy, 2nd row: injured, columns departments
x
```

```
##      [,1] [,2]
## [1,]   13   11
## [2,]    5    1
```

```r
fisher.test(x)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  x
## p-value = 0.3575
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.004593763 2.709980565
## sample estimates:
## odds ratio
##  0.2466137
```

Thus, these data do not indicate a statistically noticeable difference between the departments (P-value is 0.36, so deviation like this occurs in every third table even under the null hypothesis). Note that we cannot conclude that there is no difference between the departments, only that, with this amount of data, we do not see any statistical difference between the departments.