

Week3 Quiz

Inference Assessment 1

Load the babies dataset from babies.txt

```
babies = read.table("babies.txt", header=TRUE)
```

As Rafa has explained in previous videos, this is a very large dataset (1,236 cases), and we will pretend that it contains the entire population that we are interested in. As before, we will study the differences in birthweight between babies born to smoking and non-smoking mothers.

First, let's split this into two birthweight datasets, one of birthweights to non-smoking mothers, and the other of birthweights to smoking mothers.

```
bwt.nonsmoke = babies$bwt[babies$smoke==0]
bwt.smoke = babies$bwt[babies$smoke==1]
```

Now, we can look for the true population difference in means between smoking and non-smoking birthweights.

```
diff<-mean(bwt.nonsmoke)-mean(bwt.smoke)
sd(bwt.nonsmoke)
```

```
## [1] 17.39869
```

```
sd(bwt.smoke)
```

```
## [1] 18.09895
```

The population difference of mean birthweights is about 8.9 ounces. The standard deviations of the nonsmoking and smoking groups are about 17.4 and 18.1 ounces, respectively.

This assessment interactively reviews hypothesis testing concepts using simulations in R. We will treat the babies dataset as the full population, and draw samples from it to simulate individual experiments. We will then ask whether somebody who only received the random samples would be able to draw correct conclusions about the population.

We are interested in testing whether the birth weights of babies born to non-smoking mothers are significantly different from the birth weights of babies born to smoking mothers.

Suppose we obtain two samples, each of size N , from non-smoking mothers (dat.ns) and smoking mothers (dat.s). Following lecture, we compute the t-value, which we call tval.

```
N<-30
dat.ns<-sample(bwt.nonsmoke,N)
dat.s<-sample(bwt.smoke,N)
X.ns = mean(dat.ns)
sd.ns = sd(dat.ns)
X.s = mean(dat.s)
sd.s = sd(dat.s)
sd.diff = sqrt(sd.ns^2/N+sd.s^2/N)
tval = (X.ns - X.s)/sd.diff
```

This t-value, or t-statistic, is also returned by

```
t.test(dat.ns, dat.s)$statistic
```

```
##           t
## 4.615249
```

QUESTION 1.1

Compute the t-value (t-statistic) for the first 30 weights of non-smoking mothers and the first 30 weights of smoking mothers. (The first 30 as they appear in the bwt.nonsmoke and bwt.smoke). Confirm that the t-statistic calculated manually and by t.test() is the same. What is the t-value (t-statistic)?

```
N<-30
dat.ns<-bwt.nonsmoke[1:N]
dat.s<-bwt.smoke[1:N]
t.test(dat.ns,dat.s)$statistic
```

```
##           t
## 0.8299373
```

Recall that we summarize our data using a t-value because we know that in situations where the null hypothesis is true (what we mean when we say “under the null”) and the sample size is relatively large, this t-value will have an approximate standard normal distribution. Because we know the distribution of the t-value under the null, we can quantitatively determine how unusual the observed t-value would be if the null hypothesis were true.

The standard procedure is to examine the probability a t-value that actually does follow the null hypothesis would have larger absolute value than the absolute value of that-value we just observed – this is called a two-sided test.

One way to compute this probability is to take one minus the area under the standard normal curve between $-abs(tval)$ and $abs(tval)$. In R, we can do this using the pnorm function, which computes the area under a normal curve from negative infinity up to the value given as its first argument:

```
se <- sqrt(var(dat.ns)/length(dat.ns)+
           var(dat.s)/length(dat.s))
```

```
tstat <- diff/se
tstat
```

```
## [1] 2.160496
```

```
1-pnorm(abs(tstat)) + pnorm(-abs(tstat))
```

```
## [1] 0.03073432
```

QUESTION 1.2

Because of the symmetry of the standard normal distribution, there is a simpler way to calculate the probability that a t-value under the null could have a larger absolute value than tval. Choose the simplified calculation from the following:

- `1-2*pnorm(abs(tval))`
- `1-2*pnorm(-abs(tval))`
- `1-pnorm(-abs(tval))`
- `2*pnorm(-abs(tval))`
- `2*pnorm(abs(tval))`
- `pnorm(abs(tval))`

Answer is

```
2*pnorm(-abs(tstat))
```

```
## [1] 0.03073432
```

EXPLANATION The area outside of `-|tval|` and `|tval|` is equal to 2 times the area to the left of `-|tval|`.

No matter which way you compute it, the p-value pval is the probability that the null hypothesis could have generated a t-value more extreme than the t-value tval that we observed. If the p-value is very small, this means that observing a a value more extreme than tval would be very rare if the null hypothesis were true, and would give strong evidence that we should reject the null hypothesis. We determine how small the p-value needs to be to reject the null by deciding how often we would be willing to mistakenly reject the null hypothesis.

The standard decision rule is the following: Choose some small value alpha (in most disciplines the conventional choice is to choose $\alpha = 0.05$), and reject the null hypothesis if the p-value is less than alpha. We call alpha the significance level of the test.

It turns out if we follow this decision rule, the probability that we will reject the null hypothesis by mistake is equal to alpha. (This fact is not immediately obvious and requires some probability theory to show.) We call the event of rejecting the null hypothesis when it is in fact true a Type I error, we call the probability of making a Type I error the Type I error rate, and we say that rejecting the null hypothesis when the p-value is less than alpha controls the Type I error rate so that it is equal to alpha. (Over the course of this class, we will see a number of decision rules that we use in order to control the probabilities of other types of errors. Often we will guarantee that the probability of an error is less than some level, but in this case, we can guarantee that the probability of a Type I error is exactly equal to alpha.)

QUESTION 1.3

Which of the following sentences about a Type I error is NOT true?

- The following is another way to describe Type I error: You decided to reject the null hypothesis on the basis of data that was actually generated by the null hypothesis.
- The probability of making a Type I error is equal to the probability that the null hypothesis is true.
- The following is another way to describe Type I error: Due to random fluctuations, even though the data you observed were actually generated by the null hypothesis, the p-value calculated from the observed data was smaller than a prespecified value alpha, so you rejected it.

- In scientific practice, a Type I error constitutes reporting a “statistically significant” difference when the difference is actually equal to 0.

Answer is The probability of making a Type I error is equal to the probability that the null hypothesis is true.

Inference Assessment 2

Confidence Interval Assessment

We will continue using the dataset, babies.txt, from the previous assessment in Inference I.

```
babies = read.table("babies.txt", header=TRUE)
```

We are using the baby birthweights in the column, ‘bwt’.

Do **NOT** use the column ‘weight’, which is the mother’s weight.

You can extract the baby birthweights from smoking and non-smoking mothers with the code:

```
bwt.nonsmoke = babies$bwt[babies$smoke==0]
bwt.smoke = babies$bwt[babies$smoke==1]
```

T-TEST FOR DIFFERENCE BETWEEN MEANS (LARGE SAMPLE SIZE)

In this example, we will simulate a study where we take a sample of N=30 birthweights from each of the smoking and non-smoking groups. We will build a confidence interval for the difference between the population means using these measurements.

```
N<-30
diff<-mean(bwt.nonsmoke)-mean(bwt.smoke)
diffsample<-mean(sample(bwt.nonsmoke,N))-mean(sample(bwt.smoke,N))
```

In the last few videos, Rafa showed us how to build a confidence interval for the mean (μ) of a single group whereas in this question, we are constructing a confidence interval for a difference between groups. This is slightly more complex when sample sizes are small, but fortunately R has a built-in function that will do it for us no matter what the sample size.

The `t.test` function will take as its first two arguments two samples, and return an object that summarizes the difference between their means. This summary that has several attributes, which you can access with the `$` operator. For example, if you had samples `s1` and `s2`, you could run:

```
mytest <- t.test(s1, s2)
mytest$p.value
mytest$conf.int
```

The last two lines would return the p-value and confidence interval for the difference in means between these two samples. See the help page `?t.test` for more details.

QUESTION 1.1

Take a random sample of $N=30$ measurement from each of the smoking and nonsmoking datasets. Then compute the difference in their means and construct a 95% confidence interval for the difference using `t.test()`. Do this 1,000 times and keep track of all the confidence intervals. One way to answer this question is to use the `replicate()` function in R, which will return a 2×1000 matrix of the lower and upper bound for each of the 1000 replications.

What is the average length of the confidence interval?

```
confInt1000=replicate(1000,t.test(sample(bwt.nonsmoke,30)-sample(bwt.smoke,30))$conf.int)
mean(confInt1000[2,]-confInt1000[1,])
```

```
## [1] 18.47807
```

QUESTION 1.2

The population-level difference was

```
popdiff = mean(bwt.nonsmoke) - mean(bwt.smoke)
```

How often (what proportion of times) did the confidence intervals contain the population-level difference? That is, what proportion of times was the lower bound of the confidence interval less than popdiff and the upper bound greater than popdiff?

```
mean(confInt1000[1,<popdiff & confInt1000[2,>popdiff ]
```

```
## [1] 0.957
```

In the previous video Rafa mentioned that we should report confidence intervals whenever possible because they communicate both an effect size and the statistical significance of a result. Rafa later mentions that when comparing a difference between two groups to zero, we can tell whether the difference has a p-value of less than 0.05 based on whether the confidence interval for the difference contains zero. We will explore that statement in more detail in the next two questions.

Suppose that we have drawn samples of size $N=30$ from the non-smoking and smoking baby populations and we want to test whether the difference between their means is significant at the $\alpha=0.05$ level.

Recall that when we perform a t-test for the difference between two means, we calculate a t-value like the following.

```
dat.ns = sample(bwt.nonsmoke, 30)
dat.s = sample(bwt.smoke, 30)
X.ns = mean(dat.ns)
sd.ns = sd(dat.ns)
X.s = mean(dat.s)
sd.s = sd(dat.s)
sd.diff = sqrt(sd.ns^2/30 + sd.s^2/30)
tval = (X.ns - X.s)/sd.diff
```

Because our sample sizes are rather large, we can then use the `qnorm()` function to tell whether tval corresponds to a p-value that is less than 0.05.

```
qnorm(1-0.05/2)
```

```
## [1] 1.959964
```

This tells us that if the absolute value of `tval` is greater than 1.96, the p-value is less than 0.05 and the result is significant at the 0.05 level.

We can use the same numbers to construct a confidence interval for the difference in means between the smoking and nonsmoking populations. To do this, we follow Rafa's instruction in the CI 2 video

```
ci.upper = (X.ns-X.s) + sd.diff*1.96  
ci.lower = (X.ns-X.s) - sd.diff*1.96
```

QUESTION 1.3

Fill in the blank: the difference in means ($X_{ns} - X_s$) must have absolute value greater than _____ times `sd.diff` in order for the result to be significant (at $\alpha=0.05$).

Answer is 1.96 times

QUESTION 1.4

Fill in the blank: the difference in means ($X_{ns} - X_s$) must be a greater distance than _____ times `sd.diff` away from 0 in order for the 95% confidence interval not to contain 0.

Answer is 1.96 times

Power Calculation Assessment

In many of the analyses this week, we know that the null hypothesis is false – for example, we knew that the difference in average baby weight in the whole population is actually around 8.9. Thus, we are concerned with how often the decision rule allows us to conclude that the null hypothesis is actually false. Put another way, we would like to quantify the Type II error rate of the test, or the probability that we fail to reject the null hypothesis when the alternative hypothesis is true.

Unlike the Type I error rate, which we can characterize by assuming that the null hypothesis of “no difference” is true, the Type II error rate cannot be computed by assuming the alternative hypothesis alone because the alternative hypothesis alone does not specify a particular value for the difference, and thus does not nail down a specific distribution for the t-value under the alternative.

For this reason, when we study the Type II error rate of a hypothesis testing procedure, we need to assume a particular effect size, or hypothetical size of the difference between population means, that we wish to target. We ask questions like “What is the smallest difference I could reliably distinguish from 0 given my sample size N ?”, or more commonly, “How big does N have to be in order to detect that the absolute value of the difference is greater than zero?” Type II error control plays a major role in designing data collection procedures before you actually see the data so that you know the test you will run has enough sensitivity or power. Power is one minus the Type II error rate, or the probability that you will reject the null hypothesis when the alternative hypothesis is true.

POWER AND ALPHA

There are several aspects of a hypothesis test that affect its power for a particular effect size. Intuitively, setting a lower alpha decreases the power of the test for a given effect size because the null hypothesis will be more difficult to reject. This means that for an experiment with fixed parameters (i.e., with a predetermined sample size, recording mechanism, etc), the power of the hypothesis test trades off with its Type I error rate, no matter what effect size you target.

We can explore the tradeoff of power and Type I error concretely using the babies data. Load the babies dataset from babies.txt

```
babies = read.table("babies.txt", header=TRUE)
bwt.nonsmoke = babies$bwt[babies$smoke==0]
bwt.smoke = babies$bwt[babies$smoke==1]
```

Because we have the full population, we know what the true effect size is (about 8.93), and we can compute the power of the test for true difference between populations.

Take a random sample of $N=15$ measurements from each of the smoking (bwt.smoke) and nonsmoking (bwt.nonsmoke) groups. Then perform a t-test and compare the p-value to a significance level alpha. Do this 1,000 times. Decide whether or not to reject the null hypothesis based on three significance levels alpha=0.1, alpha = 0.05, alpha=0.01. For each experiment, keep track of whether you correctly rejected the null hypothesis at each of these significance levels (thus, each of the 1,000 experiments should produce 3 numbers to keep track of). For each significance level, in what proportion of the experiments did you correctly reject the null hypothesis?

QUESTION 3.1

What is the power at alpha=0.1?

```
N<-15
B<-1000
alpha<-0.1
rejections<-sapply(1:B,function(i){
  hf<-sample(bwt.nonsmoke,N)
  chow<-sample(bwt.smoke,N)
  t.test(hf,chow)$p.value<alpha
})
mean(rejections)
```

```
## [1] 0.414
```

QUESTION 3.2

What is the power at alpha=0.05?

```
N<-15
B<-1000
alpha<-0.05
rejections<-sapply(1:B,function(i){
  hf<-sample(bwt.nonsmoke,N)
  chow<-sample(bwt.smoke,N)
  t.test(hf,chow)$p.value<alpha
```

```
  })
  mean(rejections)
```

```
## [1] 0.26
```

QUESTION 3.3

What is the power at $\alpha=0.01$?

```
N<-15
B<-1000
alpha<-0.01
rejections<-sapply(1:B,function(i){
  hf<-sample(bwt.nonsmoke,N)
  chow<-sample(bwt.smoke,N)
  t.test(hf,chow)$p.value<alpha
})
mean(rejections)
```

```
## [1] 0.111
```

QUESTION 3.4

We will deal with the important of multiple hypothesis testing in a later course, but for now, consider this question that is very near to your heart. Suppose that one of the homework question is graded based on whether the result you reported falls within a exact 99% interval around a true value. Now suppose that 2,000 students complete the assignment, and assume that all students execute the simulation correctly. What is the expected number of student responses that would be marked wrong simply by chance?

Answer is 20. If 1% gets incorrect answer of their correct code then 1% of 2000 is 20 students.

Association Test Assessment

In the previous video, Rafa showed how to calculate a Chi-square test from a table. Here we will show how to generate the table from data which is in the form of a dataframe, so that you can then perform an association test to see if two columns have an enrichment (or depletion) of shared occurrences.

Download the assoctest.csv file into your R working directory, and then read it into R:

```
d = read.csv("assoctest.csv")
```

This dataframe reflects the allele status (either AA/Aa or aa) and the case/control status for 72 individuals.

QUESTION 2.1

Compute the Chi-square test for the association of genotype with case/control status (using the table() function and the chisq.test() function). Examine the table to see if it look enriched for association by eye. What is the X-squared statistic?


```
chisq.test(table(d))
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(d)
## X-squared = 3.3437, df = 1, p-value = 0.06746
```

QUESTION 2.2

Compute the Fisher's exact test (`fisher.test()`) for the same table. What is the p-value?

```
fisher.test(table(d))
```

```
##
## Fisher's Exact Test for Count Data
##
## data:  table(d)
## p-value = 0.05194
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.940442 8.493001
## sample estimates:
## odds ratio
##  2.758532
```

Inference Assessment 3

USING MONTE CARLO TO EXAMINE THE SAMPLE VARIANCE

In the video, Rafa showed how to examine the null distribution of the t-statistic using Monte Carlo simulations: by randomly drawing 2 sets of n samples from the non-smokers group and then calculating the t-statistic. We could then see that this distribution was approximately normal when n was large, but not when n was small (n=3).

In this assessment, we will use the same technique to examine the sample variance. A quick reminder: on the sample variance, if we have a sample 'x' with length 'n', the sample variance is defined:

$$1/(n-1) * \sum((x - \text{mean}(x))^2)$$

In R, we can simply write

```
var(x)
```

The sample variance will not typically be equal to the population (in other words, it will have some error), and the sample variance depends on the sample, which is random, so the sample variance is a random variable. Here we will examine how the sample variance is likely to be closer to the true variance as the size of the sample increases. Let's start by loading the "babies" data that was used in the previous video. The data can be downloaded from here.

```
babies = read.table("babies.txt", header=TRUE)
```

The population of nonsmoker baby weights is:

```
bwt.nonsmoke = babies$bwt[babies$smoke==0]
```

And the population variance is 302.7144:

```
pop.var = var(bwt.nonsmoke)
```

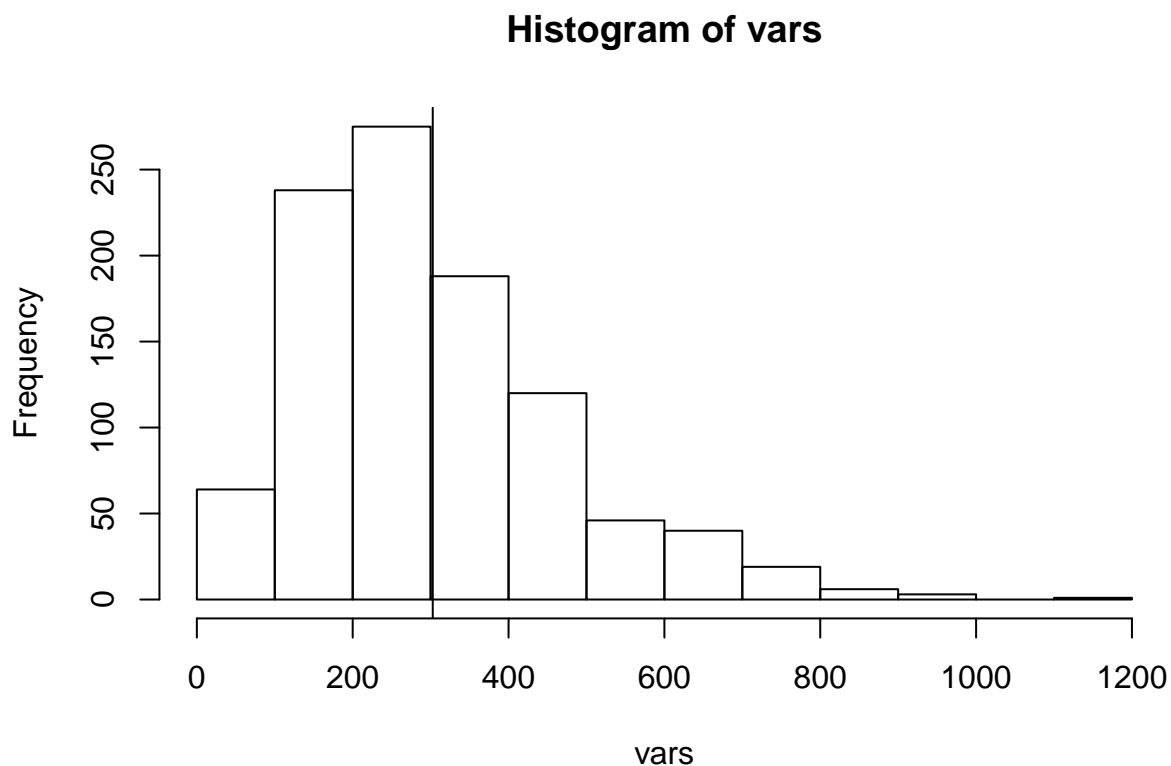
QUESTION 1.1

Use `replicate()` to perform Monte Carlo simulations. So the following line of example code will repeat the R expression, “xyz” 1000 times (you need to fill in your R code in place of “xyz”):

```
vars = replicate(1000, xyz)
```

Using Monte Carlo simulation, take 1000 samples of size 10 from `bwt.nonsmoke` and calculate the variance. Look at a histogram of `vars`, and add the population variance as a vertical line.

```
vars=replicate(1000,var(sample(bwt.nonsmoke,10)))  
hist(vars)  
abline(v=pop.var)
```



How often (what proportion of simulations) is the sample variance greater than 1.5 times the population variance? Use 1000 simulations. Your answer will change, but we have used a range of permissible answers. For the population variance, just use `pop.var` above.

```
mean(vars>(1.5*pop.var))
```

```
## [1] 0.163
```

QUESTION 1.2

Now use a sample size of 50. How often (what proportion) is the sample variance larger than 1.5 times the population variance?

```
vars=replicate(1000,var(sample(bwt.nonsmoke,50)))  
mean(vars>(1.5*pop.var))
```

```
## [1] 0.032
```

PLOT OF SAMPLE VARIANCE AND POPULATION VARIANCE

Finally, we'll make a plot to see how the sample variance estimates gets better (closer to the population variance) as the sample size increases. First, we'll make a vector of sample sizes from 2 to 400:

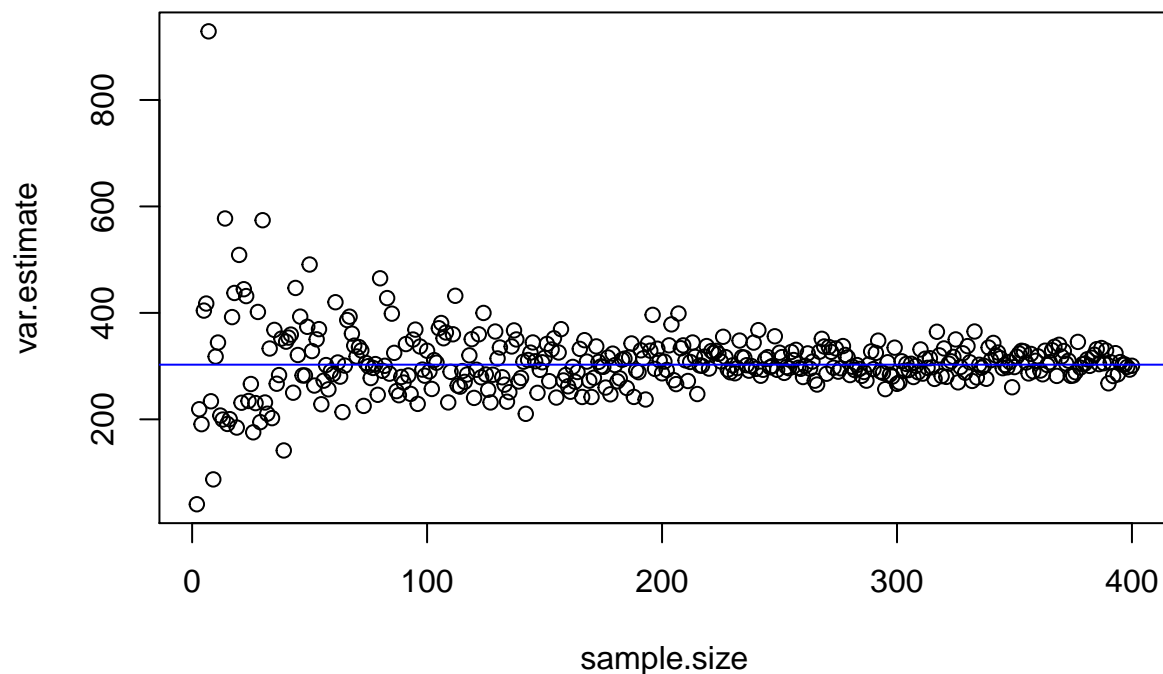
```
sample.size = 2:400
```

Now, for each sample size, take a sample from the nonsmokers of that size, and calculate the variance:

```
var.estimate = sapply(sample.size, function(n) var(sample(bwt.nonsmoke, n)))
```

Finally, plot these sample variances over their sample sizes, and draw a horizontal line of the population variance:

```
plot(sample.size, var.estimate)  
abline(h=pop.var, col="blue")
```



Permutation Assessment

In the video you just watched, Rafa provided some reasoning for why you might use a permutation test:

“Sometimes the summary statistics we form are not as simple as just taking the average or the difference in averages, and ... the null distribution is not something we can easily compute, so permutation tests give us an alternative way of doing this, of computing the null distribution.”

In the video, we computed the null distribution of the difference between the mean of the baby weights for smokers and non-smokers, using a permutation strategy. Let’s take 50 samples from each of the two groups, and suppose that we want to use a permutation test to compare these:

```
set.seed(0)
N <- 50
smokers <- sample(babies$bwt[babies$smoke==1], N)
nonsmokers <- sample(babies$bwt[babies$smoke==0], N)
```

In the video, we calculated the observed difference in means:

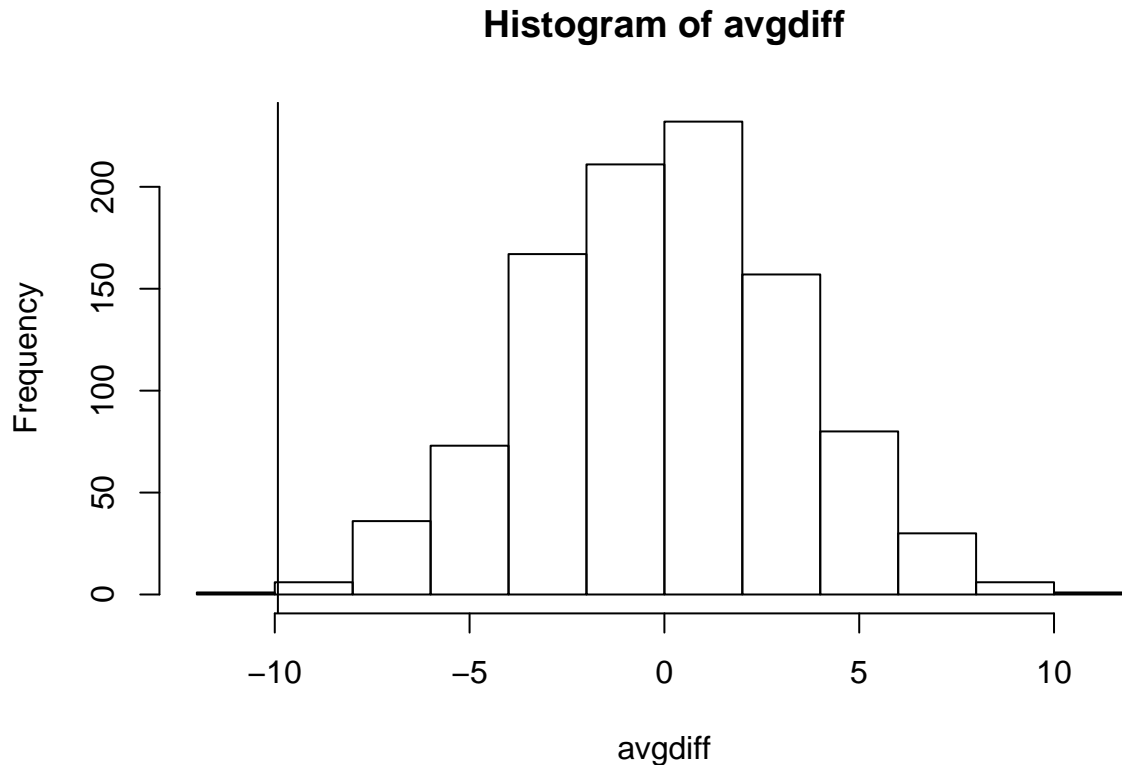
```
obs <- mean(smokers) - mean(nonsmokers)
```

Finally, we used 1000 replicated permutations to generate a null distribution. We joined the two groups, shuffled the samples, and took the first 50 in one group, and the second 50 in a second group:

```

avgdiff <- replicate(1000, {
  all <- sample(c(smokers,nonsmokers))
  smokersstar <- all[1:N]
  nonsmokersstar <- all[(N+1):(2*N)]
  return(mean(smokersstar) - mean(nonsmokersstar))
})
hist(avgdiff)
abline(v=obs)

```



Finally, we calculated a probability of seeing such a large difference between means under the null hypothesis. We use the absolute value, because we are interested if the difference was large in the positive or negative direction.

```
mean(abs(avgdiff) > abs(obs))
```

```
## [1] 0.002
```

QUESTION 2.1

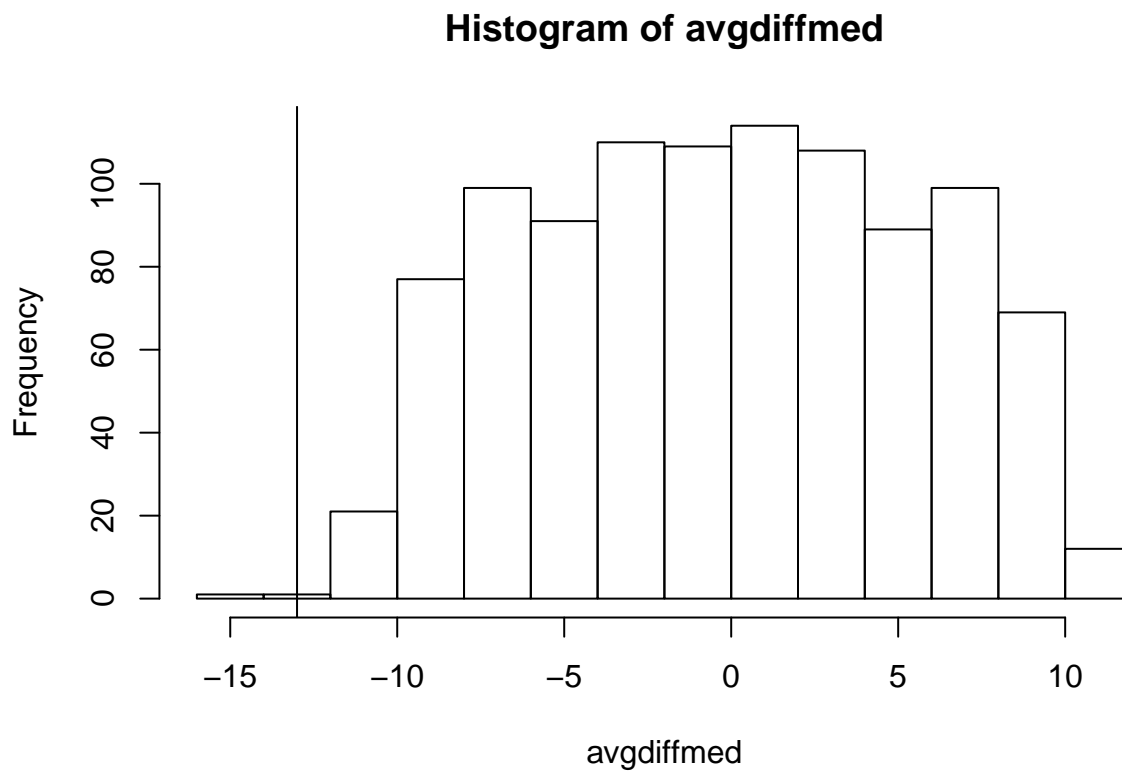
However, the difference in means (if we assume the data is normal) has some nice asymptotic results which allow us to use the t-distribution to calculate the probability of seeing so large a difference.

Suppose we are really interested in the medians of the groups, not the means. And suppose we want to know if the difference in medians is more than expected if the two groups are actually sampled from the

same population (that is, under the null hypothesis). Here is a case where the permutation test can easily be modified to give us an answer.

Use a permutation test with 1000 replications to generate a p-value for the observed difference in medians. What is the p-value for the two groups of 50 defined above?

```
set.seed(0)
N <- 50
smokers <- sample(babies$bwt[babies$smoke==1], N)
nonsmokers <- sample(babies$bwt[babies$smoke==0], N)
obsmed <- median(smokers) - median(nonsmokers)
avgdiffmed <- replicate(1000, {
  all <- sample(c(smokers,nonsmokers))
  smokersstar <- all[1:N]
  nonsmokersstar <- all[(N+1):(2*N)]
  return(median(smokersstar) - median(nonsmokersstar))
})
hist(avgdiffmed)
abline(v=obsmed)
```



```
mean(abs(avgdiffmed) > abs(obsmed))
```

```
## [1] 0.001
```