STAT2402: Week 5 Computer Laboratory

This laboratory session covers the following topics:

1. Sampling distribution of the difference between two sample proportions.
2. Hypothesis test for a difference in two population proportions.
3. Confidence intervals for the difference in two proportions.
4. Inference for odds and odds ratio.

Exercise 1: Inference for Difference of Proportions

1. The aim of this question is to demonstrate you can replicate what we have done in class for hypothesis testing and confidence intervals for the difference in two proportions. The question comes from the Statistical Sleuth, page 550, Computational exercise 9. The study relating CVD death and obesity for American Samoan Women also included men. The data for men are shown in the table below.

	CVD Death	
	Yes	No
Obese	22	1179
NonObese	22	1409

(a) Define the probability model for the number of obese Samoan men who experience CVD death.

Solution:
Let $Y_i = 1$ if the i^{th} obese Samoan man, $i = 1, \ldots, 1201$ experiences CVD death and 0 otherwise. Then $Y_i \sim \text{Bern}(\pi_Y)$ (i.e. we assume the same probability of CVD death for every obese Samoan man). If the Y_i are independent of each other, then $\sum_{i=1}^{1201} Y_i \sim \text{Binom}(1201, \pi_Y)$.

(b) Define the probability model for the number of non-obese Samoan men who experience CVD death.

Solution:
Let $Z_i = 1$ if the i^{th} non-obese Samoan man, $i = 1, \ldots, 1431$ experiences CVD death and 0 otherwise. Then $Z_i \sim \text{Bern}(\pi_Z)$ (i.e. we assume the same probability of CVD death for every non-obese Samoan man). If the Z_i are independent of each other, then $\sum_{i=1}^{1431} Z_i \sim \text{Binom}(1431, \pi_Z)$.

(c) Write the sampling distribution of the estimators of the parameters in part (a) and part (b).

Solution:
With $n_Y = 1201$ and $n_Z = 1431$, the estimators for π_Y and π_Z are:

$$\hat{\pi}_Y = \frac{\sum_{i=1}^{n_Y} Y_i}{n_Y} \quad \text{and} \quad \hat{\pi}_Z = \frac{\sum_{i=1}^{n_Z} Z_i}{n_Z}, \text{ respectively.}$$

Here the number of cases (death by CVD) and non-cases are sufficiently large for approximate Normal theory to hold. Thus, the (approximate) sampling distributions of the estimators is:

$$\hat{\pi}_Y \dot\sim N\left(\pi_Y, \frac{\pi_Y(1-\pi_Y)}{n_Y}\right) \quad \text{and} \quad \hat{\pi}_Z \dot\sim N\left(\pi_Z, \frac{\pi_Z(1-\pi_Z)}{n_Z}\right) \text{ respectively.}$$

(d) Calculate the observed sample proportions of CVD deaths for the obese and non-obese groups.

Solution:

```
n.y <- 1201
n.z <- 1431
(pi.hat.y.obs <- 22/n.y)
```

```
## [1] 0.01832

(pi.hat.z.obs <- 22/n.z)

## [1] 0.01537
```

(e) Conduct a one sided hypothesis test that obese men in American Samoa have a higher probability of CVD deaths than non-obese men. In doing so, clearly state the estimator of the difference in probabilities, the sampling distribution of this estimator, the hypothesis test, the sampling distribution under the null hypothesis (a plot of this distribution also would be nice), the p-value of the observed (difference) estimate and provide a concluding sentence.

Solution:

The estimator for the difference in probabilities is $\hat{\pi}_D = \hat{\pi}_Y - \hat{\pi}_Z$. The approximate sampling distribution of this estimator, using the normal approximation, is:

$$\hat{\pi}_D \dot{\sim} N\left(\pi_Y - \pi_Z, \frac{\pi_Y(1-\pi_Y)}{n_Y} + \frac{\pi_Z(1-\pi_Z)}{n_Z}\right)$$

We want to perform statistical inference using the hypothesis test

$$H_0 : \pi_D = 0 \text{ against } H_1 : \pi_D > 0$$

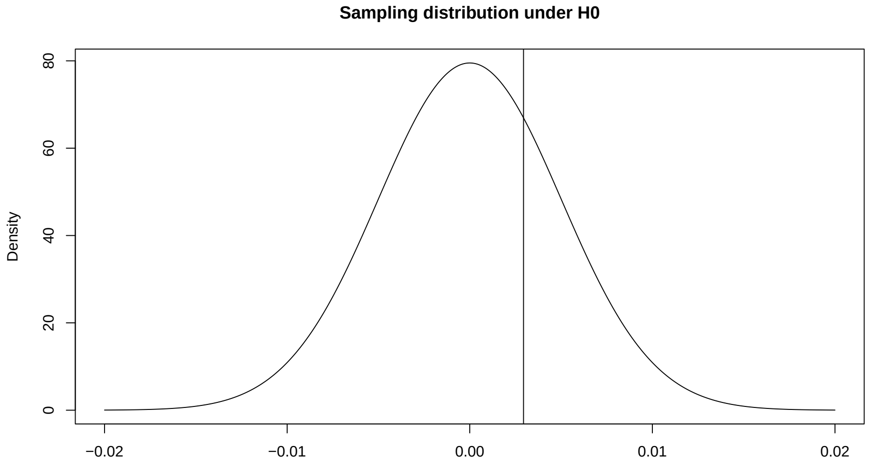
The sampling distribution of $\hat{\pi}_D$ under the null hypothesis is

$$\hat{\pi}_D \dot{\sim} N\left(0, \pi(1-\pi)\left(\frac{1}{n_Y} + \frac{1}{n_Z}\right)\right)$$

where $\pi = \pi_Y = \pi_Z$ (under H_0, $\pi_Y = \pi_Z$ and we can denote this value just by π).

We can calculate the observed difference estimate, the p-value of the observed difference estimate under H_0 and produce a plot of the approximate sampling distribution of the estimator using the following R commands:

```
(pi.hat.d.obs <- pi.hat.y.obs - pi.hat.z.obs)

## [1] 0.002944

(pi.hat.combined <- (22 + 22)/(n.y + n.z))

## [1] 0.01672

(pi.hat.se.hyp <- sqrt(pi.hat.combined * (1 - pi.hat.combined) * (1/n.y +
    1/n.z)))

## [1] 0.005017

(p.value <- pnorm(pi.hat.d.obs, 0, pi.hat.se.hyp, lower.tail = FALSE))

## [1] 0.2787

curve(dnorm(x, 0, pi.hat.se.hyp), from = -0.02, to = 0.02, n = 501,
    main = "Sampling distribution under H0", ylab = "Density", xlab = "")
abline(v = pi.hat.d.obs)
```



Given a p-value of 0.279 there is no evidence in the data against the null hypothesis. We conclude that these data provide no evidence that the probability of an obese Samoan man dying of CVD is different from the probability of a non-obese Samoan man dying of CVD.

(f) Calculate a 95% confidence interval for the difference in proportions.

Solution:
We can use the R commands:

```
pi.hat.d.se <- sqrt(pi.hat.y.obs * (1 - pi.hat.y.obs)/n.y + pi.hat.z.obs *
    (1 - pi.hat.z.obs)/n.z)
crit.val <- qnorm(1 - 0.05/2)
(CI <- pi.hat.d.obs + c(-1, 1) * crit.val * pi.hat.d.se)

## [1] -0.006963  0.012851
```

We arfe 95% confident that the true difference in proportions lies in the interval $(-0.007, 0.0129)$

2. Engineering researchers wish to compare the efficiency of truck engines using two different oil types. They randomly sampled trucks from the two oil types, Oil Type Y and Oil Type Z. Of the 112 trucks in the sample who received Oil Type Y, 84 lasted beyond 20000 oil hours. Of the 108 trucks in the sample that received Oil Type Z, 66 lasted beyond 20000 oil hours.

(a) Calculate a 95% confidence interval for the probability a truck using Oil Type Y lasts at least 20000 oil hours.

Solution:
Let Y_i be the event that truck i using Oil Type Y lasts at least 20000 oil hours. $Y_i \sim \text{Bern}(\pi_Y)$.
We know a 95% CI for π_Y is given by

$$\hat{\pi}_{Y,obs} \pm 1.96 \times \text{SE}[\hat{\pi}_Y]$$

where

$$\hat{\pi}_{Y,obs} = \frac{\sum_{i=1}^{n_Y} y_i}{n_Y} = \frac{84}{112} = 0.75$$

and

$$\text{SE}[\hat{\pi}_{Y,obs}] = \sqrt{\frac{\hat{\pi}_{Y,obs}(1 - \hat{\pi}_{Y,obs})}{n_Y}} \approx 0.0409$$

So we are 95% confident that π_Y lies in the interval $(0.6698, 0.8302)$

(b) Calculate a 95% confidence interval for the probability a truck using Oil Type Z lasts at least 20000 oil hours.

Solution:
Let Z_i be the event that truck i using Oil Type Z lasts at least 20000 oil hours. $Z_i \sim \text{Bern}(\pi_Z)$.
We know a 95% CI for π_Y is given by

$$\hat{\pi}_{Z,obs} \pm 1.96 \times \text{SE}[\hat{\pi}_Z]$$

where

$$\hat{\pi}_{Z,obs} = \frac{\sum_{i=1}^{n_Y} z_i}{n_Z} = \frac{66}{108} \approx 0.6111$$

and

$$\text{SE}[\hat{\pi}_{Z,obs}] = \sqrt{\frac{\hat{\pi}_{Z,obs}(1 - \hat{\pi}_{Z,obs})}{n_Z}} \approx 0.0469$$

So we are 95% confident that π_Y lies in the interval $(0.5192, 0.7031)$

(c) Do these confidence intervals overlap? What does this imply about the comparative effectiveness of the oil types?

Solution:
The CIs for oil type Y and Z overlap, suggesting that perhaps the two oil types have a similar effect on the longevity of trucks.

(d) Write the estimator of the difference in the two probabilities, the sampling distribution of the estimator and the standard error of the estimator?

Solution:

The estimator of the difference, $\pi_D = \pi_Y - \pi_Z$, is

$$\hat{\pi}_D = \hat{\pi}_Y - \hat{\pi}_Z$$

The sampling distribution of $\hat{\pi}_D$ is

$$\hat{\pi}_D \approx N(\pi_D, \text{Var}[\hat{\pi}_D])$$

where

$$\text{Var}[\hat{\pi}_D] = \frac{\pi_Y(1-\pi_Y)}{n_Y} + \frac{\pi_Z(1-\pi_Z)}{n_Z}$$

and

$$\text{SE}[\hat{\pi}_D] = \sqrt{\frac{\hat{\pi}_Y(1-\hat{\pi}_Y)}{n_Y} + \frac{\hat{\pi}_Z(1-\hat{\pi}_Z)}{n_Z}} \approx 0.0622$$

(e) Find a 95% confidence interval for the difference in effectiveness of the two oil types. What is your conclusions.

Solution:

The 95% CI is

$$(\hat{\pi}_{D,obs} - 1.96 \times \text{SE}[\hat{\pi}_{D,obs}], \hat{\pi}_{D,obs} + 1.96 \times \text{SE}[\hat{\pi}_{D,obs}])$$

which evaluates to be $(0.0169, 0.2609)$. This CI does not contain zero providing evidence that there is a difference in the longevity of truck engines based on the different types of oil.

(f) Why do the results in part (c) and (e) conflict? Which approach is most suitable?

Solution:

The approach in (e) is the correct CI interval to use because when looking at the difference in two independent random variables (in this case truck "longevity") the variance of the difference is the sum of the individual variances. Calculating individual CI's to compare longevity is conservative.

3. Martin and Jones (1999) investigate whether there is a systematic difference in the recollection of inanimate object orientation. University of Oxford undergraduates were asked to identify the orientation of the Hale-Bopp comet six months after the comet was visible in 1997. Students were shown 8 photographic pictures of the comet in different orientations (head of the comet facing left down, left level, left up, centre up, right up, right level, right down, or center down). The students were asked to select the correct orientation (the correct orientation was facing left down). The results are shown below:

	Correct	
	Yes	No
Left-handed	149	48
Right-handed	129	68

Table 1: Recollection of comet orientation

(a) Is there evidence of a difference in left- or right-handedness associated with correct recollection of object orientation.

Solution:

Let p_R and p_L be the proportion of correct recollection for right handed and left handed respectively. Put

$$p_D = p_R - p_L.$$

The hypotheses of interest are

$$p_D = 0 \qquad p_D \neq 0.$$

Let $\hat{p}_R$ and $\hat{p}_L$ denote the sample proportion of correct recollection for right handed left handed respectively, and put $\hat{p}_D = \hat{p}_R - \hat{p}_L$ denote the difference in sample proportions. Then

$$\hat{p}_R \overset{\cdot}{\sim} N\left(p_R, \frac{p_R(1-p_R)}{197}\right)$$
$$\hat{p}_L \overset{\cdot}{\sim} N\left(p_L, \frac{p_L(1-p_L)}{197}\right)$$

and under the null hypothesis,

$$\hat{p}_D \overset{\cdot}{\sim} N\left(0, \frac{p_R(1-p_R)}{197} + \frac{p_L(1-p_L)}{197}\right).$$

Based on the data,

$$\hat{p}_{R,Obs} = \frac{149}{197} = 0.7563$$
$$\hat{p}_{R,Obs} = \frac{129}{197} = 0.6548.$$

For the hypothesis test we need a common or pooled proportion of success,

$$\hat{p}_{C,Obs} = \frac{149+129}{197+197} = 149.3274.$$

The variance of $\hat{p}_D$ is based on the pooled proportion, so

$$\begin{aligned}\mathrm{Var}(\hat{p}_{D,Obs}) &= \frac{\hat{p}_{C,Obs}(1-p_{C,Obs)}}{197} + \frac{p_{C,Obs}(1-p_{C,Obs)}}{197}\\ &= 0.0021\end{aligned}$$

Then

$$\hat{p}_D \overset{\cdot}{\sim} N\left(0, 0.0021\right).$$

The observed difference in proportions is

$$\hat{p}_{D,Obs} = 0.7563 - 0.6548 = 0.1015.$$

The p-value of the test is

$$p - value = 2\Pr(\hat{p}_D > 0.1015) = 0.0271 < 0.05,$$

so there is sufficient evidence to reject the null hypothesis. We conclude based on this data that there is a difference between right handedness and left handedness for recollection of object orientation.

(b) Construct a 95% confidence interval for the difference.

Solution:

For the confidence interval we need the standard error of the difference in proportions, given by

$$\begin{aligned}SE &= \sqrt{\frac{\hat{p}_{R,Obs}(1-\hat{p}_{R,Obs})}{197}\frac{\hat{p}_{L,Obs}(1-\hat{p}_{L,Obs})}{197}}\\ &= 0.0461\end{aligned}$$

The a 95% confidence interval for the difference in proportions is

$$\begin{aligned}95\%CI &= (0.1015 - 1.96 \times 0.0461, 0.1015 - 1.96 \times 0.0461)\\ &= (0.0112, 0.1919).\end{aligned}$$

(c) Available in R is the function `prop.test` for inference for the difference of two proportions. Check the syntax and usage of this function.

Solution:

You will find this online. Syntax is shown in solutions to the next part.

(d) Use `prop.test` to repeat parts (a) and (b).
Solution:

```
correct <- c(149, 129)
ss <- c(197, 197)
prop.test(x = correct, n = ss, alternative = "two.sided")
```

```
## 
##  2-sample test for equality of proportions with continuity
##  correction
## 
## data:  correct out of ss
## X-squared = 4.4, df = 1, p-value = 0.04
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.006998 0.196048
## sample estimates:
## prop 1 prop 2 
## 0.7563 0.6548 
```

Note that the confidence interval is different from that calculated in part (b). If you look at the output in this part, you will notice the line
`2-sample test for equality of proportions with continuity correction`
This makes the normal approximation to the binomial more accurate.

Exercise 2: Inference for Odds Ratios

1. In the STAT2402 LMS site contains a data set called "`heart.txt`". Download this to your appropriate working directory. Read and inspect the data in `R` by typing in the console window:

```
heart <- read.table(file = "../Data/heart.txt", header = TRUE)
summary(heart)
```

```
##      death           anterior        hcabg            center    
##  Min.   :0.0000   Min.   :0.0    Min.   :0.0000   Min.   :1255  
##  1st Qu.:0.0000   1st Qu.:0.0    1st Qu.:0.0000   1st Qu.:3318  
##  Median :0.0000   Median :0.0    Median :0.0000   Median :5029  
##  Mean   :0.0449   Mean   :0.5    Mean   :0.0336   Mean   :5043  
##  3rd Qu.:0.0000   3rd Qu.:1.0    3rd Qu.:0.0000   3rd Qu.:6331  
##  Max.   :1.0000   Max.   :1.0    Max.   :1.0000   Max.   :9668  
##                   NA's   :692                                   
##      killip         agegrp          age       
##  Min.   :1.00   Min.   :1.00   Min.   : 40.0  
##  1st Qu.:1.00   1st Qu.:1.00   1st Qu.: 54.0  
##  Median :1.00   Median :2.00   Median : 66.0  
##  Mean   :1.35   Mean   :2.16   Mean   : 65.8  
##  3rd Qu.:2.00   3rd Qu.:3.00   3rd Qu.: 76.0  
##  Max.   :4.00   Max.   :4.00   Max.   :100.0  
##  NA's   :235                                  
```

The data set contains, among other variables,

- `Death`: 1 = die; 0 = survive.
- `anterior`: 1 = anterior ; 0 = inferior.

The variable `anterior` records whether a person has had a myocardial infarction (injury) in the anterior (front) of the heart or in the inferior (back) of the heart. The variable *death* records whether the patient died within 48 hours of admission into hospital with the infarction.

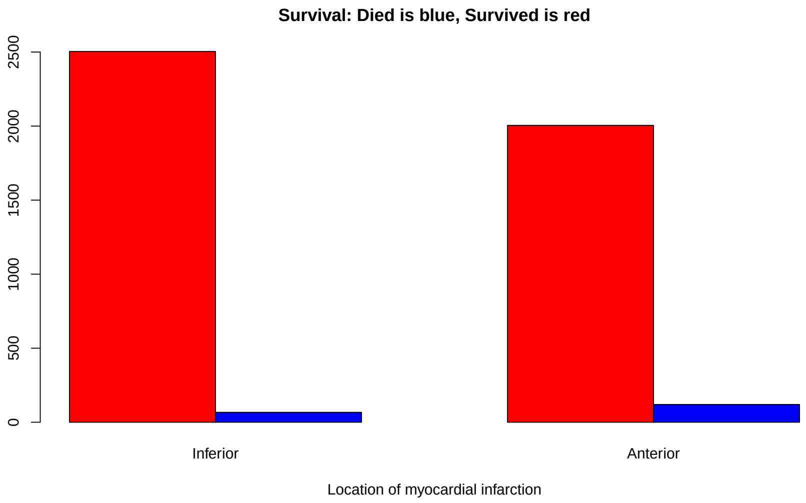
(a) Tabulate death versus anterior by

```
(heart.table <- table(data.frame(heart$death, heart$anterior)))
```

```
##            heart.anterior
## heart.death    0    1
##           0 2504 2005
##           1   67  120
```

and display the data using the function “barplot” (described in the R notes).
Solution:

```
barplot(heart.table, xlab = "Location of myocardial infarction", col = c("red",
    "blue"), main = "Survival: Died is blue, Survived is red", beside = TRUE,
    names.arg = c("Inferior", "Anterior"))
```



(b) Calculate the observed odds ratio of death from anterior infarction to inferior infarction. The following commands show you how to do so directly from the table you just created:

```
(odds.ratio <- heart.table[1, 1] * heart.table[2, 2]/(heart.table[1,
    2] * heart.table[2, 1]))
```

```
## [1] 2.237
```

Now calculate the logarithm of the odds ratio using

```
(log.odds.ratio <- log(odds.ratio))
```

```
## [1] 0.805
```

(c) Calculate the approximate standard error of $\log(\hat{\phi})$ for a confidence interval using the command:

```
(log.phi.hat.se.ci <- sqrt(sum(1/heart.table)))
```

```
## [1] 0.1554
```

Why does this command give the standard error of the log-odds—refer to the slides of Week 3 (short cut method). Can the same command be used to calculate the standard error of $\log(\hat{\phi})$ for a hypothesis test?
Solution:
In an hypothesis test this standard error would commonly *not* be used as the standard error will be determined by value of ϕ stipulated under the null hypothesis.

(d) Now calculate the confidence interval (symmetric on the log scale) and then back transform to the *original* scale by:

```
UL <- exp(log.odds.ratio + 1.96 * log.phi.hat.se.ci)
LL <- exp(log.odds.ratio - 1.96 * log.phi.hat.se.ci)
c(LL, UL)
```

```
## [1] 1.649 3.033
```

Does the interval suggest the odds will be greater than 1?

Solution:
Yes, we are 95% confident that the true odds ratio is in this interval, and the interval does not include 1 and all values in that interval are greater than 1.

2. The aim of this question is to demonstrate you understand and are able to conduct inference on the difference of two odds. This question is again from the Statistical Sleuth, page 568, Computational exercise 9. The same data in Exercise 1) part a) above is used.

The `R` code for the calculations is given below.

```
n.y <- 1201
n.z <- 1431
pi.hat.y <- 22/n.y
pi.hat.z <- 22/n.z
pi.hat <- (22 + 22)/(n.y + n.z)
odds.hat.y <- pi.hat.y/(1 - pi.hat.y)
odds.hat.z <- pi.hat.z/(1 - pi.hat.z)
phi.hat <- odds.hat.y/odds.hat.z
phi.hat.se.hyp <- sqrt(1/(n.y * pi.hat * (1 - pi.hat)) + 1/(n.z *
    pi.hat * (1 - pi.hat)))
phi.hat.se <- sqrt(1/22 + 1/22 + 1/1179 + 1/1409)
```

(a) Write down the estimator of the odds of CVD death for obese American Samoan men.

Solution:
Let Y_i be the event of the i^{th} obese American Samoan man dies of CVD. Then $n_Y = 1201$, $Y_i \sim \text{Bern}(\pi_Y)$ and the odds of CVD death for obese American Samoan men are $\omega_Y = \frac{\pi_Y}{1-\pi_Y}$.
The estimator for π_Y is $\hat{\pi}_Y = \frac{1}{n_Y}\sum_{i=1}^{n_Y} Y_i$ and the estimator for ω_Y is $\hat{\omega}_Y = \frac{\hat{\pi}_Y}{1-\hat{\pi}_Y}$.

(b) Write down the estimator of the odds of CVD death for non-obese American Samoan men.

Solution:
Let Z_i be the event of the i^{th} non-obese American Samoan man dies of CVD. Then $n_Z = 1431$, $Z_i \sim \text{Bern}(\pi_Z)$ and the odds of CVD death for non-obese American Samoan men are $\omega_Z = \frac{\pi_Z}{1-\pi_Z}$.
The estimator for π_Z is $\hat{\pi}_Z = \frac{1}{n_Z}\sum_{i=1}^{n_Z} Z_i$ and the estimator for ω_Z is $\hat{\omega}_Z = \frac{\hat{\pi}_Z}{1-\hat{\pi}_Z}$.

(c) Write down the estimator of the odds ratio of CVD death for obese American Samoan men to non-obese American Samoan men.

Solution:
The odds ratio of CVD death for obese American Samoan men to non-obsese American Samoan men is $\phi = \frac{\omega_Y}{\omega_Z}$. An estimator for this odds ratio is $\hat{\phi} = \frac{\hat{\omega}_Y}{\hat{\omega}_Z}$.

(d) Write the sampling distribution of odds ratio estimator—or an appropriate transformation of this estimator.

Solution:
If the entries in the table are large enough, then $\log(\hat{\phi})$ has an approximate normal distribution:

$$\log(\hat{\phi}) \dot{\sim} N\left(\log(\phi), \text{Var}[\log(\hat{\phi})]\right)$$

where

$$\text{Var}[\log(\hat{\phi})] = \frac{1}{n_Y \pi_Y (1-\pi_Y)} + \frac{1}{n_Z \pi_Z (1-\pi_Z)}$$

(e) Conduct a hypothesis test for the equality of the above odds. In doing so, clearly state the sampling distribution under the null hypothesis (a plot of this distribution also would be nice), the p-value of our observed (log odds ratio) estimate and provide a concluding sentence.

Solution:
Let us use a two-sided alternative, i.e. we want to test

$$H_0 : \omega_Y = \omega_Z \text{ against } H_1 : \omega_Y \neq \omega_Z$$

Or, equivalently

$$H_0 : \phi = 1 \text{ against } H_1 : \phi \neq 1$$

Or, equivalently

$$H_0 : \log(\phi) = 0 \text{ against } H_1 : \log(\phi) \neq 0$$

Calculating the standard error under the null hypothesis yields:

$$\text{SE}[\log(\hat{\phi})] = \sqrt{\frac{1}{n_Y \hat{\pi}_{obs}(1 - \hat{\pi}_{obs})} + \frac{1}{n_z \hat{\pi}_{obs}(1 - \hat{\pi}_{obs})}} \approx 0.3052$$

where $\hat{\pi}_{obs} = (22 + 22)/(1201 + 1431) = 0.0167$ is the estimate of the total proportion of men who died from CVD.

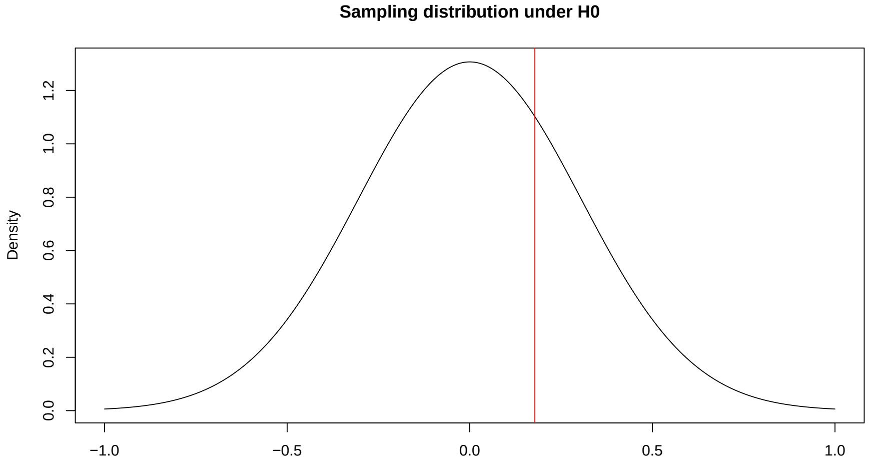
The observed log-odds ratio is $\log(\hat{\phi}_{obs}) = 0.1782$ so the p-value is

$$\text{p-value} = 2 \times \Pr[\log(\hat{\phi}) > 0.17821 | \log(\phi) = 0] = 0.5593$$

This large p-value leads to retention of the H_0. The data provides no evidence that the true odds ratio of CVD death for obese American Samoan men to to non-obese American Samoan men is not one.

A plot of the approximate sampling distribution of $\log(\hat{\phi})$ can be produced using the following commands:

```r
n.y <- 1201
n.z <- 1431
pi.hat <- (22 + 22)/(n.y + n.z)
phi.hat <- 22 * 1409/(22 * 1179)
phi.hat.se.hyp <- sqrt(1/(n.y * pi.hat * (1 - pi.hat)) + 1/(n.z *
    pi.hat * (1 - pi.hat)))
curve(dnorm(x, 0, phi.hat.se.hyp), from = -1, to = 1, n = 501, ylab = "Density",
    xlab = "", main = "Sampling distribution under H0")
abline(v = log(phi.hat), col = "red")
```



The observed value $\log(\hat{\phi}_{obs})$ is indicated by the red vertical line.

(f) Calculate a 95% confidence interval for the odds ratio.

Solution:

Calculating the standard error using the short-cut formula:

$$\text{SE}[\log(\hat{\phi})] = \sqrt{\frac{1}{22} + \frac{1}{22} + \frac{1}{1179} + \frac{1}{1409}} \approx 0.3041$$

We can calculate the 95% CI on the log scale as

$$\log(\hat{\phi}_{obs}) \pm 1.96 \times \text{SE}[\log(\hat{\phi})] = (-0.4178, 0.7742)$$

and taking the exponential of the upper and lower limits to put the CI back on the original scale we have $(0.6585, 2.1689)$. We are 95% confident that this interval contains the true odds-ratio of CVD death for obease American Samoan men to non-obese American Samoan men.

	Cancer	Control	Totals
Smokers	83	72	155
NonSmokers	3	14	17
Totals	86	86	172

3. The following data are from Chapter 18, Statistical Sleuth (p 552), and relates smoking to Lung Cancer. The original data are described in Dorn (1954).

   These data are retrospective and samples were taken from each level of the response (lung cancer) and the explanatory variable (smoking) was determined for each subject in these samples. Such studies cannot be used to estimate the individual proportions of smokers and non-smokers who get lung cancer (or the differences). (Why not?). The odds ratio however is the same regardless of which variable is considered the response (Why?).

   For the above table:

   (a) What is the observed odds of getting cancer if you smoke? Provide an interpretation of these odds.
   Solution:
   The R code for the calculations is given below.

```
n1 <- 155
n2 <- 17
pi.hat.1 <- 83/n1
pi.hat.2 <- 3/n2
pi.hat <- (83 + 3)/(n1 + n2)
odds.hat.1 <- pi.hat.1/(1 - pi.hat.1)
odds.hat.2 <- pi.hat.2/(1 - pi.hat.2)
phi.hat <- odds.hat.1/odds.hat.2
log.phi.hat = log(phi.hat)
se.hyp <- sqrt(1/(n1 * pi.hat * (1 - pi.hat)) + 1/(n2 * pi.hat * (1 -
    pi.hat)))
se <- sqrt(1/83 + 1/72 + 1/3 + 1/14)
U <- log.phi.hat + 1.96 * se
L <- log.oh.hat - 1.96 * se

## Error in eval(expr, envir, enclos):  object 'log.oh.hat' not found
```

   Let Y_i be the event of the i^{th} individual who smokes getting cancer. Then $n_Y = 155$ and $Y_i \sim \text{Bern}(\pi_Y)$. The observed estimate for π_Y is $\hat{\pi}_Y = \frac{83}{155} \approx 0.0183$ and the observed odds estimate is

$$\hat{\omega}_{Y,obs} = \frac{\hat{\pi}_{Y,obs}}{1 - \hat{\pi}_{Y,obs}} \approx 1.1528$$

   If you are a smoker you are 1.1528 times more likely to get cancer than not.

   (b) What is the observed odds of getting cancer if you do not smoke? Provide an interpretation of these odds.
   Solution:
   Let Z_i be the event of the i^{th} individual who does not smoke getting cancer. Then $n_Z = 17$ and $Z_i \sim \text{Bern}(\pi_Z)$. The observed estimate for π_Z is $\hat{\pi}_Z = \frac{3}{17} \approx 0.1765$ and the observed odds estimate is

$$\hat{\omega}_{Z,obs} = \frac{\hat{\pi}_{Z,obs}}{1 - \hat{\pi}_{Z,obs}} \approx 0.2143$$

   If you are a non-smoker you are 0.2143 times more likely to get cancer than not.

   (c) What is the observed odds ratio of cancer for smokers to cancer for non-smokers? What is the interpretation of this odds ratio?
   Solution:
   The observed odds ratio of cancer for smokers to cancer for non-smokers is

$$\hat{\phi}_{obs} = \frac{\hat{\omega}_{Y,obs}}{\hat{\omega}_{Z,obs}} \approx 5.38$$

   The odds of a smoker getting cancer are about 5.38 times the odds of a non-smoker getting cancer.

(d) Test whether the odds of cancer are greater for smokers than non-smokers. Be clear in your reasoning and provide a concluding sentence.

Solution:
Let ω_Y denotes the odds that a smoker gets cancer, ω_Z the odds that a non-smoker get cancer and $\phi = \frac{\omega_Y}{\omega_Z}$ the odds ratio of cancer for smokers to cancer for non-smokers. We want to test

$$H_0 : \omega_Y = \omega_Z \text{ against } H_1 : \omega_Y > \omega_Z$$

Or, equivalently

$$H_0 : \phi = 1 \text{ against } H_1 : \phi > 1$$

Or, equivalently

$$H_0 : \log(\phi) = 0 \text{ against } H_1 : \log(\phi) > 0$$

Assume the natural logarithm of the estimator of the odds ratio, has the approximate Gaussian distribution:

$$\log(\hat{\phi}) \stackrel{\cdot}{\sim} N\left(\log(\phi), \text{Var}[\log(\hat{\phi})]\right)$$

where

$$\text{Var}[\log(\hat{\phi})] = \frac{1}{n_Y \pi_Y (1-\pi_Y)} + \frac{1}{n_Z \pi_Z (1-\pi_Z)}$$

Calculating the standard error under the null hypothesis yields:

$$\text{SE}[\log(\hat{\phi})] = \sqrt{\frac{1}{n_Y \hat{\pi}_{obs}(1-\hat{\pi}_{obs})} + \frac{1}{n_z \hat{\pi}_{obs}(1-\hat{\pi}_{obs})}} \approx 0.511$$

where $\hat{\pi}_{obs} = (83+3)/(155+17) = 0.5$ is the estimate of the total proportion of people who got cancer.

The observed log-odds ratio is $\log(\hat{\phi}_{obs}) = 1.6826$ so the p-value is

$$\text{p-value} = \Pr[\log(\hat{\phi}) > 1.6826 | \log(\phi) = 0] = 4.958 \times 10^{-4}$$

The small p-value leads to rejection of the H_0 at conventional significance levels (5% or 1%). The data provides evidence that the true odds ratio of getting cancer for smokers compared to non-smokers is not one. Given the point estimate for the odds ratio, there is evidence in the data to conclude that the odds of getting cancer is higher for smokers than for non-smokers.

(e) Provide a 95% confidence interval for the odds ratio of cancer for smokers to non-smokers. Be clear in your reasoning and provide a concluding sentence.

Solution:
Calculating the standard error using the short-cut formula:

$$\text{SE}[\log(\hat{\phi})] = \sqrt{\frac{1}{83} + \frac{1}{72} + \frac{1}{3} + \frac{1}{14}} \approx 0.6563$$

We can calculate the 95% CI on the log scale as

$$\log(\hat{\phi}_{obs}) \pm 1.96 \times \text{SE}[\log(\hat{\phi})] = (0.0112, 2.969)$$

and taking the exponential of the upper and lower limits to put the CI back on the original scale we have $(2.964, 9.7632)$. We are 95% confident that this interval contains the true odds-ratio of cancer for smokers to non-smokers.

References

1) Dorn, HF (1954). “The relationship of Cancer of the Lung and the Use of Tobacco”, *American Statistician* **8**, 7–13.