140.652 Problem Set 6 Solutions

Problem 1

Consider the hypothesis testing problem of comparing two binomial probabilities $H_0 : p_1 = p_2$. Show that the square of statistic $(\hat{p}_1 - \hat{p}_2)/\text{SE}_{\hat{p}_1 - \hat{p}_2}$ is the same as the χ^2 statistic. Here, the standard error in the denominator is calculated under the null hypothesis. (Clearly define any notation you introduce.)

Suppose we have a standard 2×2 table given below:

	Treatment	No Treatment	
Symptoms	n_{11}	n_{12}	$n_1 = n_{1+}$
No Symptoms	n_{21}	n_{22}	$n_2 = n_{2+}$
	n_{+1}	n_{+2}	

Now, define

$$\hat{p}_1 = \frac{n_{11}}{n_1}$$
$$\hat{p}_2 = \frac{n_{21}}{n_2}$$
$$\hat{p} = \frac{n_{11} + n_{21}}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Additionally, recall that

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

The χ^2 statistic is given by,

$$\begin{aligned}
\chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\
&= \frac{(n_{11} - n_1\hat{p})^2}{n_1\hat{p}} + \frac{(n_{12} - n_1(1-\hat{p}))^2}{n_1(1-\hat{p})} + \frac{(n_{21} - n_2\hat{p})^2}{n_2\hat{p}} + \frac{(n_{22} - n_2(1-\hat{p}))^2}{n_2(1-\hat{p})} \\
&= \frac{(n_1\hat{p}_1 - n_1\hat{p})^2}{n_1\hat{p}} + \frac{(n_1(1-\hat{p}_1) - n_1(1-\hat{p}))^2}{n_1(1-\hat{p})} + \frac{(n_2\hat{p}_2 - n_2\hat{p})^2}{n_2\hat{p}} + \frac{(n_2(1-\hat{p}_2) - n_2(1-\hat{p}))^2}{n_2(1-\hat{p})} \\
&= \frac{n_1^2(\hat{p}_1 - \hat{p})^2}{n_1\hat{p}} - \frac{n_1^2(\hat{p}_1 + \hat{p})^2}{n_1(1-\hat{p})} + \frac{n_2^2(\hat{p}_2 - \hat{p})^2}{n_2\hat{p}} + \frac{n_2^2(\hat{p}_2 - \hat{p})^2}{n_2(1-\hat{p})} \\
&= \frac{n_1(\hat{p}_1 - \hat{p})^2}{\hat{p}} + \frac{n_1(\hat{p}_1 - \hat{p})^2}{(1-\hat{p})} + \frac{n_2(\hat{p}_2 - \hat{p})^2}{\hat{p}} + \frac{n_2(\hat{p}_2 - \hat{p})^2}{(1-\hat{p})} \\
&= \frac{n_1(1-\hat{p})(\hat{p}_1 - \hat{p})^2 + n_1\hat{p}(\hat{p}_1 - \hat{p})^2 + n_2(1-\hat{p})(\hat{p}_2 - \hat{p})^2 + n_2\hat{p}(\hat{p}_2 - \hat{p})^2}{\hat{p}(1-\hat{p})} \\
&= \frac{n_1(\hat{p}_1 - \hat{p})^2 + n_2(\hat{p}_2 - \hat{p})^2}{\hat{p}(1-\hat{p})} \\
&= \frac{n_1\hat{p}_1^2 + n_2\hat{p}_2^2 - 2(n_1\hat{p}_1 + n_2\hat{p}_2)\hat{p} + (n_1 + n_2)\hat{p}^2}{\hat{p}(1-\hat{p})} \\
&= \frac{n_1\hat{p}_1^2 + n_2\hat{p}_2^2 - 2(n_1 + n_2)\hat{p}^2 + (n_1 + n_2)\hat{p}^2}{\hat{p}(1-\hat{p})} \\
&= \frac{n_1\hat{p}_1^2 + n_2\hat{p}_2^2 - (n_1 + n_2)\hat{p}^2}{\hat{p}(1-\hat{p})}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{n_1\hat{p}_1^2 + n_2\hat{p}_2^2 - (n_1+n_2)\frac{(n_1\hat{p}_1+n_2\hat{p}_2)^2}{(n_1+n_2)^2}}{\hat{p}(1-\hat{p})} \\
&= \frac{n_1\hat{p}_1^2 + n_2\hat{p}_2^2 - \frac{(n_1\hat{p}_1+n_2\hat{p}_2)^2}{(n_1+n_2)}}{\hat{p}(1-\hat{p})} \\
&= \frac{\frac{1}{n_1+n_2}\left[n_1(n_1+n_2)\hat{p}_1^2 + n_2(n_1+n_2)\hat{p}_2^2 - (n_1\hat{p}_1+n_2\hat{p}_2)^2\right]}{\hat{p}(1-\hat{p})} \\
&= \frac{\frac{1}{n_1+n_2}\left[n_1^2\hat{p}_1^2 + n_1n_2(\hat{p}_1^2+\hat{p}_2^2) + n_2^2\hat{p}_2^2 - (n_1^2p_1^2 + 2n_1n_2\hat{p}_1\hat{p}_2 + n_2^2\hat{p}_2^2)\right]}{\hat{p}(1-\hat{p})} \\
&= \frac{\frac{n_1n_2}{n_1+n_2}(\hat{p}_1^2 - 2\hat{p}_1\hat{p}_2 + \hat{p}_2^2)}{(n_1+n_2)\hat{p}(1-\hat{p})} = \frac{(\hat{p}-\hat{p}_2)^2}{\frac{n_1+n_2}{n_1n_2}\hat{p}(1-\hat{p})} \\
&= \frac{(\hat{p}-\hat{p}_2)^2}{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\hat{p}(1-\hat{p})} = \left(\frac{\hat{p}_1-\hat{p}_2}{SE_{\hat{p}_1-\hat{p}_2}}\right)^2
\end{aligned}
$$

Thus, the χ^2 statistic is equal to the statistic $\frac{\hat{p}_1-\hat{p}_2}{SE_{\hat{p}_1-\hat{p}_2}}$ squared.

Problem 2

A study of the effectiveness of *streptokinase* in the treatment of patients who have been hospitalized after myocardial infarction involves a treated and control group. In the streptokinase group, 20 of 150 patients died within 12 months. In the -control group, 40 of 190 died with 12 months.

a. Test equivalence of the two proportions.

Let p_T and p_C denote the proportion of patients who passed away in the treatment and control groups, respectively. We want to test,

$$H_0 : p_T = p_C \qquad H_A \qquad : p_T \neq p_C$$

To test the null hypothesis, we can use the score test. Here $\hat{p}_T = \frac{20}{150}$, $\hat{p}_C = \frac{40}{190}$, and $\hat{p} = \frac{\hat{p}_T n_T + \hat{p}_C n_C}{n_T + n_C}$. The test statistic is given by,

$$TS = \frac{\hat{p}_C - \hat{p}_T}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_T}+\frac{1}{n_C}\right)}} = \frac{\frac{40}{190} - \frac{20}{150}}{\sqrt{\frac{20+40}{150+190}\left(1-\frac{20+40}{150+190}\right)\left(\frac{1}{150}+\frac{1}{190}\right)}} = 1.85$$

For large n_T and n_C the score test statistic follows a standard normal, implying that $P(|TS| > 1.85)$ is,

```
2* pnorm(1.85, lower.tail = FALSE)
```

```
[1] 0.06431355
```

Thus, with $\alpha = 0.05$, we fail to reject the null and conclude that the proportions may not be significantly different between the two groups.

b. Give confidence intervals for the absolute change in proportions, the relative risk and odds ratio.

Confidence interval for the absolute change in proportions:

Since the score test is not easily invertible and the Wald CI performs poorly in practice, we will use the Agresti/Caffo interval for the confidence interval for the absolute change in proportions. The Agresti/Caffo interval is given by,

$$(\tilde{p}_T - \tilde{p}_C) \pm z_{1-\alpha/2}\sqrt{\frac{\tilde{p}_C(1-\tilde{p}_C)}{\tilde{n}_C} + \frac{\tilde{p}_T(1-\tilde{p}_T)}{\tilde{n}_T}}$$

where,

$$\begin{aligned}\tilde{p}_C &= \frac{40+1}{190+2}\\ \tilde{p}_T &= \frac{20+1}{150+2}\\ \tilde{n}_C &= 190+2\\ \tilde{n}_T &= 150+2\end{aligned}$$

```
pC_tilde <- (40 + 1)/(190 + 2)
pT_tilde <- (20 + 1)/(150 + 2)
nC_tilde <- 190 + 2
nT_tilde <- 150 + 2

(pT_tilde - pC_tilde) + c(-1, 1) * qnorm(0.975) *
  sqrt(pC_tilde*(1 - pC_tilde)/nC_tilde + pT_tilde*(1 - pT_tilde)/nT_tilde)
```

```
[1] -0.155191935  0.004424391
```

Confidence interval for the relative risk:

To get the confidence interval for the relative risk, we can get the confidence interval for the log of the relative risk then exponentiate. Recall that the confidence interval for the log relative risk is given by,

$$\log\left(\frac{\hat{p}_T}{\hat{p}_C}\right) \pm z_{1-\alpha/2}\sqrt{\frac{1-\hat{p}_T}{\hat{p}_T n_T} + \frac{1-\hat{p}_C}{\hat{p}_C n_C}}$$

```
pT_hat <- 20/150
pC_hat <- 40/190
nT <- 150
nC <- 190

exp(log(pT_hat/pC_hat) + c(-1, 1) * qnorm(0.975) *
      sqrt((1-pT_hat)/(pT_hat*nT) + (1-pC_hat)/(pC_hat*nC)))
```

```
[1] 0.3871359 1.0360989
```

Confidence interval for the odds ratio:

Similarly, we can derive a confidence interval for the odds ratio by exponentiating the confidence interval for the log odds ratio. Recall that the confidence interval for the log odds ratio is given by,

$$\log\left(\frac{\hat{p}_T/(1-\hat{p}_T)}{\hat{p}_C/(1-\hat{p}_C)}\right) \pm z_{1-\alpha/2}\sqrt{\frac{1}{\hat{p}_T n_T} + \frac{1}{(1-\hat{p}_T)n_T} + \frac{1}{\hat{p}_C n_C} + \frac{1}{(1-\hat{p}_C)n_C}}$$

```
exp(log((pT_hat/(1-pT_hat))/(pC_hat/(1-pC_hat))) + c(-1, 1) * qnorm(0.975) *
      sqrt(1/(pT_hat*nT) + 1/((1-pT_hat)*nT) + 1/(pC_hat*nC) + 1/((1-pC_hat)*nC)))
```

```
[1] 0.3211208 1.0364953
```

c. Create Bayesian credible intervals for the risk difference, risk ratio and odds ratio. Plot the posterior for each and interpret the results.

We can use simulations to derive Bayesian credible intervals. Let us assume that,

- For the treatment group, $p_T \sim Beta(1,1)$ and $X_T \sim Binom(n_T, p_T)$ where X_T is the number of individuals who died in the treatment group.
- For the control group, $p_C \sim Beta(1,1)$ and $X_C \sim Binom(n_C, p_C)$ where X_C is the number of individuals who died in the control group.

As we have shown earlier, the Beta distribution is a conjugate prior for the binomial distribution. Thus,

- The posterior distribution for the treatment group is given by $p_T|X_T \sim Beta(X_T + 1, n_T - X_T + 1)$
- The posterior distribution for the control group is given by $p_X|X_C \sim Beta(X_C + 1, n_C - X_C + 1)$

```
# Sample pT, pC from the posterior distribution 1000 times
pT <- rbeta(10000, 20 + 1, 150 - 20 + 1)
pC <- rbeta(10000, 40 + 1, 190 - 40 + 1)

# Calculate the absolute risk difference and look at the quantiles from 0.025 to 0.975
risk_diff <- pT - pC
quantile(risk_diff, c(0.025, 0.975))
```

```
       2.5%        97.5%
-0.153891083  0.003784263
```

```
# Calculate the relative risk difference and look at the quantiles from 0.025 to 0.975
relative_risk <- pT/pC
quantile(relative_risk, c(0.025, 0.975))
```

```
     2.5%     97.5%
0.3886344 1.0212485
```

```
# Calculate the odds ratio and look at the quantiles from 0.025 to 0.975
odds_ratio <- (pT/(1-pT))/(pC/(1-pC))
quantile(odds_ratio, c(0.025, 0.975))
```

```
    2.5%    97.5%
0.323323 1.026133
```

Problem 3

Researchers are interested in estimating the natural log of the proportion of people in the population with hypertension. In a random sample of n subjects, let X be the number with hypertension. Derive a confidence interval for the natural log of the proportion of people with hypertension. Assume that n is large.

Define $\hat{p} = \frac{X}{n}$. By the CLT, $\sqrt{n}(\hat{p} - p) \xrightarrow{D} N(0, p(1-p))$. Since n is large, we can apply the Delta method with the function $f(x) = \log(x)$, yielding

$$\sqrt{n}(f(\hat{p}) - f(p)) \xrightarrow{D} N\left(0, \frac{p(1-p)}{p^2}\right)$$

Thus, the $(1-\alpha)\%$ CI is given by,

$$\log(\hat{p}) \pm z_{1-\alpha/2}\sqrt{\frac{1-p}{np}} \approx \log(\hat{p}) \pm z_{1-\alpha/2}\sqrt{\frac{1-\hat{p}}{n\hat{p}}}$$

Problem 4

This problem considers the delta method.

a. Derive the asymptotic standard error for $\sqrt{\hat{p}}$ where $\hat{p}$ is a binomial sample proportion.

Suppose we have $X_1, X_2, \cdot, X_n$ iid Bernoulli(p) random variables and define $\hat{p} = \bar{X}$. Since X_i are bounded, their variance is finite. Applying the CLT, we have that

$$\sqrt{n}(\hat{p} - p) \xrightarrow{D} N(0, p(1-p))$$

Let $f(x) = \sqrt{x}$. Then $f'(x) = -\frac{1}{2\sqrt{x}}$. By the Delta method,

$$\sqrt{n}(\sqrt{\hat{p}} - \sqrt{p}) \xrightarrow{D} N\left(0, \frac{p(1-p)}{4pn}\right) = N\left(0, \frac{1-p}{4n}\right)$$

We can approximate the variance using $\frac{1-\hat{p}}{4n}$.

b. Assume that $n = 200$ and $p = .5$. Implement a simulation study to verify that the delta method results in approximately normally distributed variables.

Our simulation is as follows:

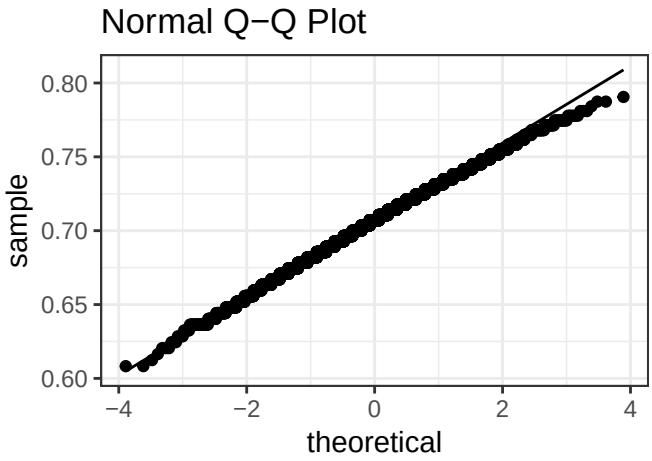
1. Simulate 10,000 Binomial(200, 0.5) random variables. This represents $10,000 \times 200$ Bernoulli(0.5) trials.
2. Calculate the estimated proportion $\hat{p}$ by dividing the 10,000 Binomial random variables by 200.
3. Transform the data by taking the square root of $\hat{p}$, and calculate the empirical mean and variance.

We can then compare the empirical mean and variance to those obtained by from the Delta method in part (a). Using a qq-plot and looking at the density, we can see that our simulation yields approximately results as the theoretical results derived from the Delta method.

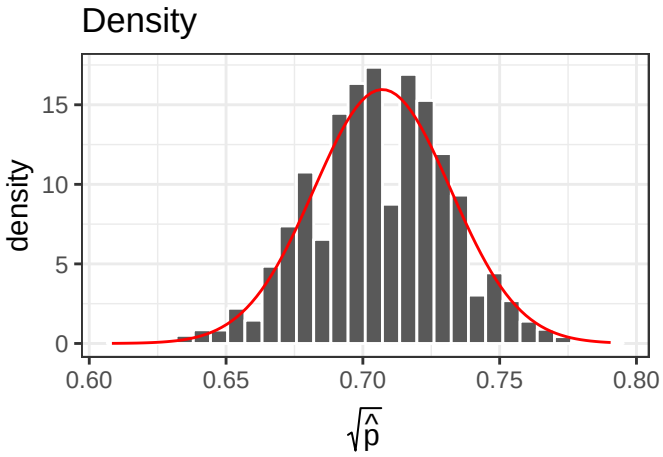
```
binom <- rbinom(10000, 200, 0.5)  # Simulate 10000 binomial random variables
phat  <- binom/200                # Calculate proportions

results <- data.frame(empirical = c(mean(sqrt(phat)), var(sqrt(phat))),
           delta = c(sqrt(0.5), (1-0.5)/(4*200)))
rownames(results) <- c("mean", "variance")
results
```

```
           empirical     delta
mean     0.7064447556 0.7071068
variance 0.0006218695 0.0006250
```

```
# Check normality of the data
data.frame(sqrt_phat = sqrt(phat)) %>%
  ggplot(aes(sample = sqrt_phat)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Normal Q-Q Plot") +
  theme_bw()
```




```
# Look at density plot
data.frame(sqrt_phat = sqrt(phat)) %>%
  ggplot(aes(x = sqrt_phat)) +
  geom_histogram(aes(y = ..density..), bins = 30, color = 'white') +
  stat_function(color = 'red',
                fun = function(x)
                  dnorm(x, mean = sqrt(0.5), sd = sqrt((1-0.5)/(4*200)))) +
  labs(title = "Density", x = TeX('$\\sqrt{\\hat{p}}$')) +
  theme_bw()
```



Problem 5

In this homework, we will evaluate the performance of the log odds ratio interval estimate

$$\log \hat{OR} \pm 1.96\sqrt{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}}.$$

Use R to generate 1,000 random binomials with `n1` trials and `p1` success probability; call this vector `x`. Use R to generate 1,000 random binomials with `n2` trials and `p2` success probability; call this vector `y`. Squash these to vectors together with the command `z = cbind(x, y)`. Now, create 1,000 sample odds ratios with the command:

```
OR = apply(z, 1,
           function(x) x[1] * (n2 - x[2]) / (x[2] * (n1 - x[1]))
          )
```

Log these odds ratios to obtain 1,000 sample log odds ratios. Now obtain 1,000 standard errors with the command

```
SELOGOR = apply(z, 1,
           function(x) sqrt(1 / x[1] + 1 / (n1 - x[1]) +
                            1 / x[2] + 1 / (n2 - x[2])
                           )
          )
```

Now, see how often the interval for the log odds ratio contains the true log odds ratio. Repeat this process for all of the following combinations

```
p1 = .1; p2 = .1; n1 = 100; n2 = 100
p1 = .1; p2 = .5; n1 = 100; n2 = 100
p1 = .1; p2 = .9; n1 = 100; n2 = 100
p1 = .5; p2 = .5; n1 = 100; n2 = 100
p1 = .5; p2 = .9; n1 = 100; n2 = 100
p1 = .9; p2 = .9; n1 = 100; n2 = 100
```

Summarize your findings.

```
# Function to calculate the proportion of time the CI for the log odds ratio (LOR)
# contains the true log odds ratio
# Input:
#     p1      success probability for first binomial sample
#     p2      success probability for the second binomial sample
#     n1      size of first binomial sample
#     n2      size of second binomial sample
# Output:
#     prop    proportion of CI that contain the true log odds ratio
logodds_prop <- function(p1, p2, n1, n2){

  # Calculate true LOR
  trueLOR <- log((p1/(1 - p1))/(p2/(1 - p2)))

  # Simulate random binomials, merge two binomials into matrix z
  x <- rbinom(1000, size = n1, prob = p1)
  y <- rbinom(1000, size = n2, prob = p2)
  z <- cbind(x, y)

  # Get 1000 sample OR and log sample OR
  OR <- apply(z, 1, function(x) x[1] * (n2 - x[2]) / (x[2] * (n1 - x[1])))
  LOR <- log(OR)

  # Get standard errors for log ORs
  SELOGOR = apply(z, 1, function(x)
    sqrt(1 / x[1] + 1 / (n1 - x[1]) +  1 / x[2] + 1 / (n2 - x[2])))

  # Create CI
  CI_lower <- LOR - 1.96*SELOGOR
  CI_upper <- LOR + 1.96*SELOGOR

  # Calculate proportion of times the true LOR is contained within the CI
  prop <- mean(CI_lower <= trueLOR & CI_upper >= trueLOR)
  return(prop)
}
```

```r
set.seed(12345)
# Create matrix of parameters
p1 <- c(0.1, 0.1, 0.1, 0.5, 0.5, 0.9)
p2 <- c(0.1, 0.5, 0.9, 0.5, 0.9, 0.9)
n1 <- rep(100, 6)
n2 <- rep(100, 6)
param <- cbind(p1, p2, n1, n2)

# Get proportion of times the true LOR is contained within the CI using apply
apply(param, 1, function(x) logodds_prop(x[1], x[2], x[3], x[4]))
```

```
[1] 0.963 0.959 0.955 0.943 0.950 0.957
```

We see that our 95% confidence interval for the log odds ratio contains the true log odds approximately 95% of the time, as expected.

Problem 6

In this homework, we will also evaluate the performance of the log relative risk interval estimate

$$\log \hat{RR} \pm 1.96\sqrt{(1-\hat{p}_1)/(\hat{p}_1 n_1) + (1-\hat{p}_2)/(\hat{p}_2 n_2)}$$

Use R to generate $1,000$ random binomials with `n1` trials and `p1` success probability; call this vector `x`. Use R to generate $1,000$ random binomials with `n2` trials and `p2` success probability; call this vector `y`. Squash these to vectors together with the command `z = cbind(x, y)`. Now, create $1,000$ sample risk ratios with the command

```r
RR = apply(z, 1,
           function(x) (x[1] / n1) / (x[2] / n2)
          )
```

Log these risk ratios to obtain $1,000$ sample log risk ratios. Now obtain $1,000$ standard errors with the command

```r
SELOGRR = apply(z, 1,
             function(x) {
                phat1 <- x[1] / n1
                phat2 <- x[2] / n2
                sqrt((1 - phat1) / phat1 / n1 + (1 - phat2) / phat2 / n2)
             }
          )
```

Now, see how often the interval for the log relative risk contains the true log relative risk. Repeat this process for the following combinations

```r
p1 = .1; p2 = .1; n1 = 100; n2 = 100
p1 = .1; p2 = .5; n1 = 100; n2 = 100
p1 = .1; p2 = .9; n1 = 100; n2 = 100
p1 = .5; p2 = .5; n1 = 100; n2 = 100
p1 = .5; p2 = .9; n1 = 100; n2 = 100
p1 = .9; p2 = .9; n1 = 100; n2 = 100
```

Summarize your findings.

```r
# Function to calculate the proportion of time the CI for the log relative risk (LRR)
# contains the true log relative risk
# Input:
#     p1      success probability for first binomial sample
#     p2      success probability for the second binomial sample
#     n1      size of first binomial sample
#     n2      size of second binomial sample
# Output:
#     prop    proportion of CI that contain the true log relative risk
logrr_prop <- function(p1, p2, n1, n2){

  # Calculate true LRR
  trueLRR <- log(p1/p2)

  # Simulate random binomials, merge two binomials into matrix z
  x <- rbinom(1000, size = n1, prob = p1)
  y <- rbinom(1000, size = n2, prob = p2)
  z <- cbind(x, y)

  # Get 1000 sample RR and log sample RR
  RR <- apply(z, 1, function(x) (x[1] / n1) / (x[2] / n2))
  LRR <- log(RR)

  # Get standard errors for log RRs
  SELOGRR = apply(z, 1, function(x) {
    phat1 <- x[1] / n1
    phat2 <- x[2] / n2
    sqrt((1 - phat1) / phat1 / n1 + (1 - phat2) / phat2 / n2)}
  )

  # Create CI
  CI_lower <- LRR - 1.96*SELOGRR
  CI_upper <- LRR + 1.96*SELOGRR

  # Calculate proportion of times the true LRR is contained within the CI
  prop <- mean(CI_lower <= trueLRR & CI_upper >= trueLRR)
  return(prop)
}

set.seed(12345)
# Create matrix of parameters
p1 <- c(0.1, 0.1, 0.1, 0.5, 0.5, 0.9)
p2 <- c(0.1, 0.5, 0.9, 0.5, 0.9, 0.9)
n1 <- rep(100, 6)
n2 <- rep(100, 6)
param <- cbind(p1, p2, n1, n2)

# Get proportion of times the true LRR is contained within the CI using apply
apply(param, 1, function(x) logrr_prop(x[1], x[2], x[3], x[4]))
```

```
[1] 0.965 0.967 0.952 0.959 0.954 0.956
```

We see that our 95% confidence interval for the log relative risk contains the true relative risk approximately 95% of the time, as expected.

Problem 7

The following data show the results of caries surveys in five towns and also the fluoride content of the drinking water.

Area Fluoride p.p.m.	Surrey and Essex 0.15	Slough 0.9	Harwick 2.0	Burnham 3.5	West Meresa 5.8	Total
Number children with with caries	243	83	60	31	39	456
Number children with caries free teeth	16	36	32	31	12	127
Number examined	259	119	92	62	51	583

The data refer to samples of children aged 12-14 only.

For all parts of the question, assume

- Children in these five towns were selected randomly (independence)
- The probability of caries for children in the same town are the same (identically distributed)

Additionally, for each sub-question, the null hypothesis is that the two relevant probabilities are the same, while the alternative hypothesis is that they are different. We use a test of level $\alpha = 0.05$.

a. Estimate the odds ratio for the probability of caries for lowest and highest (.15 to 5.8) categories of fluoride. Interpret these results.

```
p1 <- 243/259
p2 <- 39/51
n1 <- 259
n2 <- 51

# Calculate LOR and SE of LOR
LOR <- log((p1/(1-p1))/(p2/(1-p2)))
SE_logOR <- sqrt(1/(n1*p1) + 1/(n1*(1-p1)) + 1/(n2*p2) + 1/(n2*(1-p2)))

# Get 95% CI of LOR
CI_LOR <- LOR + c(-1, 1)*qnorm(0.975)*SE_logOR

# Estimate OR and 95\% CI for OR
exp(LOR)
```

```
[1] 4.673077
```

```
exp(CI_LOR)
```

```
[1]  2.055514 10.623935
```

The estimate for the odds ratio for the probability of caries for lowest and highest (.15 to 5.8) categories of fluoride is 4.67 with a 95% CI of (2.055, 10.624). This means that the odds of caries for children at Surrey and Essex is 4.67 times the odds of caries for children at West Meresa. Since the 95% CI does not contain 1, we conclude that there may be a difference in odds for the probability of carries between these two towns.

b. Estimate the relative risk for the probability of caries for lowest and highest (.15 to 5.8) categories of fluoride. Interpret these results.

```
# Calculate LRR and SE of LRR
LRR <- log(p1/p2)
```

```r
SE_logRR <- sqrt((1-p1)/(n1*p1) + (1-p2)/(n2*p2))

# Get 95% CI of LOR
CI_LRR <- LRR + c(-1, 1)*qnorm(0.975)*SE_logRR

# Estimate OR and 95\% CI for OR
exp(LRR)
```

```
[1] 1.226908
```

```r
exp(CI_LRR)
```

```
[1] 1.050310 1.433199
```

The estimate for the relative risk for the probability of caries for lowest and highest (.15 to 5.8) categories of fluoride is 1.23 with a 95% CI of (1.05, 1.43). This means that the risk of caries for children at Surrey and Essex is 1.23 times the risk of caries for children at West Meresa. Since the 95% CI does not contain 1, we conclude that there may be a difference in risk for the probability of carries between these two towns.

c. Estimate the risk difference for the probability of caries for lowest and highest (.15 to 5.8) categories of fluoride. Interpret these results.

```r
# Calculate risk difference and SE of risk difference
RD <- p1 - p2
RD
```

```
[1] 0.1735181
```

```r
SE_RD <- sqrt(p1 * (1 - p1)/n1 + p2 * (1 - p2)/n2)

# Get 95% CI of RD
RD + c(-1, 1)*qnorm(0.975)*SE_RD
```

```
[1] 0.05346586 0.29357025
```

The estimate for the risk difference for the probability of caries for lowest and highest (.15 to 5.8) categories of fluoride is 0.17 with a 95% CI of (0.05, 0.29). This means that to risk of caries for children at Surrey and Essex is 0.17 more than the risk of caries for children at West Meresa. Since the 95% CI does not contain 0, we conclude that there may be a difference in risk for the probability of carries between these two towns.

d. Test the equivalence for the probability of caries between the lowest and highest fluoride concentrations (.15 to 5.8) give a P-value and interpret your results.

To test the equivalence in probabilities of caries between the lowest and highest fluoride concentrations, we can use a score test as n_1 and n_2 are large.

```r
# Calculate Test Statistic
p <- (243 + 39)/(259 + 51)
ts <- (p1 - p2)/sqrt(p * (1 - p) * (1/n1 + 1/n2))

# Calculate p-value with a two-sided alternative hypothesis
2*pnorm(ts, lower.tail = FALSE)
```

```
[1] 7.767783e-05
```

With $\alpha = 0.05$, we reject the null and conclude that there may be a difference in the probabilities of caries between the lowest and highest fluoride concentrations.

Problem 8

A case-control study of esophageal cancer was performed. Daily alcohol consumption was ascertained (80+ gm = high, 0 − 79 gm = low). The data was stratified by 3 age groups.

	Alcohol			Alcohol			Alcohol	
	H	L		H	L		H	L
case	8	5	case	25	21	case	50	61
control	52	164	control	29	138	control	27	208
	Age 35-44			Age 45-54			Age 55-64	

Give a confidence interval estimate of the odds ratio in the Oldest age group.

Because the number of subject is sufficiently large, we can appeal to asymptotics (e.g. CLT and Delta method) to derive a confidence interval estimate of the odds ratio in th oldest age group. Recall that the 95% CI for the log odds ratio is given by,

$$\log\left(\frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)}\right) \pm 1.96\sqrt{1/n_{11}+1/n_{12}+1/n_{21}+1/n_{22}}$$

To get the 95% CI for the odds ratio, we exponentiate the 95% CI for the log odds ratio.

```
n11 <- 50
n12 <- 61
n21 <- 27
n22 <- 208

# Estimate Log odds ratio and SE of log odds ratio
LOR <- log(n11/n12/(n21/n22))
SE_LOR <- sqrt(1/n11 + 1/n12 + 1/n21 + 1/n22)

# 95\% CI for LOR
CI_LOR <- LOR + c(-1, 1) * qnorm(0.975) * SE_LOR

# 95\% CI for OR
exp(CI_LOR)
```

```
[1]  3.649635 10.925217
```

The 95% CI for the odds ratio for the oldest age group is (3.65, 10.93).

Problem 9

In a study of aquaporins, 120 frog eggs were randomized, 60 to receive a protein treatment and 60 controls. If the treatment of the protein is effective, the frog eggs would implode. The resulting data was

	Imploded	Did not	Total
Treated	50	10	60
Control	20	40	60
Totals	70	50	120

State the appropriate hypotheses and report and interpret a P-value.

We want to see whether there was a difference in p_T and p_C, the probabilities of implosion for eggs in the treatment and control groups, respectively. That is, we want to test

$$H_0 : p_T = p_C$$
$$H_A : p_T \neq p_C$$

Assume that all eggs are independent and identically distributed within treatment groups. Then, we can use a score test of a difference in proportions to test our hypothesis.

```
pT <- 50/60
pC <- 20/60

# Calculate score test statistic
p <- (50+20)/(60 + 60)
TS <- (pT - pC)/sqrt(p*(1-p)*(1/60 + 1/60))

# Calculate p-value
2*pnorm(TS, lower.tail = FALSE)
```

```
[1] 2.77738e-08
```

Thus, with $\alpha = 0.05$ and a p-value of 2.78×10^{-8}, we reject the null and conclude that there may be a difference in implosion probabilities between the treatment and control groups.

Problem 10

Let $\hat{p}$ be the sample proportion from a binomial experiment with n trials. Recall that the standard error of $\hat{p}$ is $\sqrt{\hat{p}(1-\hat{p})/n}$. Define $f(\hat{p}) = \log\{\hat{p}/(1-\hat{p})\}$ as the sample log odds. Note, the following fact might be useful:

$$f'(x) = \frac{1}{x(1-x)}$$

Use the delta method to create a confidence interval for the sample log odds.

Suppose we have $X_1, X_2, \cdots, X_n$ iid Bernoulli(p) trials. By the CLT, we know that

$$\sqrt{n}(\hat{p} - p) \xrightarrow{D} N\left(0, p(1-p)\right)$$

Define $f(x) = \log\left(\frac{x}{1-x}\right)$. Then, by the Delta method, we have

$$\sqrt{n}(f(\hat{p}) - f(p)) \xrightarrow{D} N\left(0, f'(p)^2 p(1-p)\right) = N\left(0, \frac{p(1-p)}{p^2(1-p)^2}\right) = N\left(0, \frac{1}{p(1-p)}\right)$$

Equivalently,

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) - \log\left(\frac{p}{1-p}\right) \xrightarrow{D} N\left(0, \frac{1}{np(1-p)}\right)$$

Using the sample proportion to estimate the true proportion, the $(1-\alpha)\%$ confidence interval for the sample log-odds is given by,

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) \pm z_{1-\alpha/2}\sqrt{\frac{1}{n\hat{p}(1-\hat{p})}}$$

Problem 11

Refer to the previous problem. Let $\hat{p}_1$ be the sample proportion from one binomial experiment with n_1 trials and $\hat{p}_2$ be the sample proportion from a second with n_2 trials. Define the log odds ratio to be $f(\hat{p}_1) - f(\hat{p}_2)$. Use your answer to part 2 to derive a confidence interval for the log odds ratio.

The log odds ratio is given by,

$$\log\left(\frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)}\right) = \log\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) - \log\left(\frac{\hat{p}_2}{1-\hat{p}_2}\right)$$

From problem 10, we know that,

$$\log\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) - \log\left(\frac{p_1}{1-p_1}\right) \xrightarrow{D} N\left(0, \frac{1}{n_1 p_1(1-p_1)}\right)$$
$$\log\left(\frac{\hat{p}_2}{1-\hat{p}_2}\right) - \log\left(\frac{p_2}{1-p_2}\right) \xrightarrow{D} N\left(0, \frac{1}{n_2 p_2(1-p_2)}\right)$$

Thus, assuming independence of the 2 experiments, we have

$$\log\left(\frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)}\right) - \log\left(\frac{p_1/(1-p_1)}{p_2(1-p_2)}\right) \xrightarrow{D} N\left(0, \frac{1}{n_1 p_1(1-p_1)} + \frac{1}{n_2 p_2(1-p_2)}\right)$$

Using the sample proportions to estimate the true proportions, the $(1-\alpha)\%$ confidence interval for the sample log-odds ratio is given by,

$$\log\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) - \log\left(\frac{\hat{p}_2}{1-\hat{p}_2}\right) \pm z_{1-\alpha/2}\sqrt{\frac{1}{n_1 p_1(1-p_1)} + \frac{1}{n_2 p_2(1-p_2)}}$$

Or equivalently,

$$\log\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) - \log\left(\frac{\hat{p}_2}{1-\hat{p}_2}\right) \pm z_{1-\alpha/2}\sqrt{\frac{1}{n_1 p_1} + \frac{1}{n_1(1-p_1)} + \frac{1}{n_2 p_2} + \frac{1}{n_2(1-p_2)}}$$

Problem 12

Two drugs, A and B , are being investigated in a randomized trial with the data are given below. Investigators would like to know if the Drug A has a greater probability of side effects than drug B.

	None	Side effects	N
Drug A	10	30	40
Drug B	30	10	40

a. State relevant null and alternative hypotheses and perform the relevant test.

We want to see whether there was a difference in p_A and p_B, the probabilities of side effects for drug A and drug B, respectively. That is, we want to test

$$H_0 : p_A = p_B$$
$$H_A : p_A \neq p_B$$

Assume that all patients are independent and identically distributed within drug groups. Then, we can use a score test of a difference in proportions to test our hypothesis.

```r
pA <- 10/40
pB <- 30/40
nA <- nB <- 40

# Calculate score test statistic
p <- (10 + 30)/(40 + 40)
TS <- (pA - pB)/sqrt(p*(1-p)*(1/nA + 1/nB))

# Calculate p-value
2*pnorm(TS)
```

```
[1] 7.744216e-06
```

Thus, with $\alpha = 0.05$ and a p-value of 7.74×10^{-6}, we reject the null and conclude that there may be a difference in the probabilities of side effects between drug A and drug B.

b. Estimate a confidence interval for the odds ratio, relative risk and risk difference. Interpret these results.

```r
# Get CI for odds ratio by exponentiating CI for log odds ratio
LOR <- log((pA/(1-pA))/(pB/(1-pB)))
SE_logOR <- sqrt(1/(nA*pA) + 1/(nA*(1-pA)) + 1/(nB*pB) + 1/(nB*(1-pB)))

CI_LOR  <- LOR + c(-1, 1)*qnorm(0.975)*SE_logOR # Get 95% CI of LOR
OR      <- c(exp(LOR), exp(CI_LOR))             # OR estimate and 95% CI of OR

# Get CI for relative risk by exponentiating CI for log relative risk
LRR <- log(pA/pB)
SE_logRR <- sqrt((1-pA)/(nA*pA) + (1-pB)/(nB*pB))

CI_LRR  <- LRR + c(-1, 1)*qnorm(0.975)*SE_logRR # Get 95% CI of LOR
RR      <- c(exp(LRR), exp(CI_LRR))             # RR estimate and 95% CI of RR

# Get CI for risk difference and SE of risk difference
RD <- pA - pB
p <- (pA*nA + pB*nB)/(nA + nB)
SE_RD <- sqrt(p * (1 - p)*(1/nA + 1/nB))
RD <- c(RD, RD + c(-1, 1)*qnorm(0.975)*SE_RD)

# Organize all the CIs in a convenient dataframe
CI <- as.data.frame(rbind(OR, RR, RD),
                    row.names = c("OddsRatio", "RelativeRisk", "RiskDifference"))
colnames(CI) <- c("Estimate", "2.5%", "97.5%")
CI
```

```
                 Estimate        2.5%      97.5%
OddsRatio       0.1111111  0.04038303  0.3057145
RelativeRisk    0.3333333  0.18930323  0.5869478
RiskDifference -0.5000000 -0.71913064 -0.2808694
```

Problem 13

You are flipping a coin and would like to test if it is fair. You flip it 10 times and get 8 heads. Specify relevant hypotheses and report and interpret an exact P-value.

Assume that the coin flips are independent and identically distributed and that the probability of flipping heads is p. We are interested in testing,

$$H_0 : p = 0.5$$
$$H_A : p \neq 0.5$$

Since the sample size is small ($n = 10$), I decided to test my hypothesis using an exact test. Under the null, the probability of getting as or more extreme value is the probability of flipping 8 or more heads or 2 or fewer heads. This probability is given by,

$$P(X \geq 8 \text{ or } X \leq 2) = \sum_{i \leq 2 \text{ or } i \geq 8}^{10} \binom{10}{i} 0.5^i (1 - 0.5)^{10-i}$$

```
pbinom(7, 10, 0.5, lower.tail = FALSE) + pbinom(2, 10, 0.5)
```

```
[1] 0.109375
```

For $\alpha = 0.05$ and a p-value of 0.11, we fail to reject the null and conclude that the coin may be fair.