۱-

$H_0: p_{us} = 0.38$

$H_A: p_{us} \neq 0.38$

$\hat{p}_{observed} = 0.17 \quad , n = 2254 \quad , \quad , SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.38 \times 0.62}{2254}} = 0.0102$

$Z = \frac{\hat{p}_{observed} - p_{us}}{SE} = \frac{0.17 - 0.38}{0.0102} = -20.58$

$p - value = P(|Z| > 20.58) \approx 0 < \alpha = 0.05 \quad \rightarrow \quad \boldsymbol{Reject\ H_0}$

پس می‌توان گفت درصد آمریکایی‌هایی که از گوشی خود برای دسترسی به اینترنت استفاده می‌کنند، با درصد چینی‌هایی که از گوشی خود برای دسترسی به اینترنت استفاده می‌کنند، یکسان نمی‌باشد.

در اینجا p_value به این معنی است که احتمال مشاهده‌ی proportion = 17% برای آمریکایی‌ها با فرض صحیح بودن H0 (با فرض اینکه درصد چینی‌ها و آمریکایی‌های استفاده کننده از گوشی برای دسترسی به اینترنت یکسان باشد) چقدر است. مقدار p_value بدست آمده تقریباً صفر است که باعث می‌شود بتوانیم نتیجه بگیریم 17% مشاهده شده نمی‌تواند یه مقدار تصادفی باشد و اختلاف آماری معناداری بین درصد چینی‌ها و آمریکایی‌ها وجود دارد.

حال برای محاسبه‌ی confidence Interval داریم:

$confidence\ Interval = point\ estimate \pm Z^* SE \quad , \qquad point\ estimate = observed$

$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.17 \times 0.83}{2254}} \approx 0.008 , Confidence\ level = 95\% \rightarrow Z^* = 1.96$

$0.17 \pm 1.96 \times 0.008 = 0.17 \pm 0.01568 \quad \rightarrow \quad (0.15432, 0.18568)$

$0.38 \notin (0.15432, 0.18568) \rightarrow \boldsymbol{Reject\ H_0}$

ما ۹۵٪ مطمئن هستیم که ۱۵.۴٪ تا ۱۸.۵٪ آمریکایی‌ها از گوشی خود برای دسترسی به اینترنت استفاده می‌کنند.

۲- برای حل این مسئله می‌توانیم از استنباط با استفاده از simulation کمک بگیریم یا به طور تئوری با استفاده از فرمول توزیع binomial نیز استفاده کنیم.(چون تعدادی simulation نداریم از فرمول استفاده می‌کنیم)

- از یک سکه سالم استفاده می‌کنیم(شیر: انتخاب صحیح)
- P-value: احتمال اینکه در ۸۰ بار پرتاب سکه حداقل ۵۳ بار شیر بیاید.

$H_0: p = 0.5$

$H_A: p > 0.5$

$$p - value = P(X \geq 53 | \, p = 0.5) = \sum_{k=53}^{80} \binom{80}{k} (0.5)^k (0.5)^{80-k}$$

برای محاسبه احتمال بالا با استفاده از R داریم:

```
> sum(dbinom(53:80,80,0.5))
[1] 0.002434077
```

$p - value = 0.002 < \alpha = 0.05 \quad \rightarrow \quad \textbf{Reject } \boldsymbol{H_0}$

توضیح قسمت *a* و *b* سوال باهم:

یعنی احتمال مشاهده‌ی چنین داده ای با فرض صحیح بودن H_0 (تصادفی بودن انتخاب شرکت کنندگان) خیلی کم است. پس H_0 را رد می‌کنیم. یعنی تشخیص نوع نوشابه توسط شرکت کنندگان بهتر از انتخاب تصادفی است. به عبارت دیگر می‌توان گفت شرکت کنندگان می‌توانند نوع نوشابه را تشخیص دهند(از آنجایی که *p-value* بدست آمده هم دارای مقدار خیلی کوچکی است، این جمله را می‌توان با اطمینان بیشتری گفت)

p-value در اینجا به این معنی است که احتمال مشاهده‌ی چنین proportionی ($\frac{53}{80}$) یا شدیدتر با فرض صحیح بودن null hypothesis (با فرض تصادفی بودن انتخاب نوع نوشابه توسط شرکت‌کنندگان) چقدره.

اگر با استفاده از simulation بخواهیم معنی p-value را در اینجا توصیف کنیم، می‌توان گفت: درصد simulation هایی که در آنها proportion تشخیص های درست $\frac{53}{80}$ یا بزرگتر از آن است.

۳-

a)

	Nevaripine	Lopinavir	total
virologic failure	26	10	36
Non-failure	94	110	204
total	120	120	240
$\boldsymbol{\widehat{P}}$	0.217	0.083	0.15

---

b) شرایط CLT برای مقایسه‌ی دوتا proportion را بررسی می‌کنیم:

- استقلال درون گروهی و بین گروهی داریم.
- شرط success-failure برقرار نیست:
  - Nevaripine: $26 \times 0.15 = 3.9 < 10 \;\; and \; 26 \times 0.85 = 22.1 \;\geq 10$
  - Lopinavir: $10 \times 0.15 = 1.5 < 10 \;\; and \; 10 \times 0.85 = 8.5 \; < 10$

فرض می‌کنیم شرایط CLT برقرار است.

$\hat{p}_{pool} = 0.15$

$H_0: p_{\text{Nevaripine}} - p_{\text{Lopinavir}} = 0$

$H_A: p_{\text{Nevaripine}} - p_{\text{Lopinavir}} \neq 0$

$$SE = \sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_{\text{Nevaripine}}} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_{\text{Lopinavir}}}} = \sqrt{\frac{0.15 \times 0.85}{120} + \frac{0.15 \times 0.85}{120}} \approx 0.046$$

$\hat{p}_{\text{Nevaripine}} - \hat{p}_{\text{Lopinavir}} \sim N(mean = 0, SE = 0.046)$

$point\ estimate: \hat{p}_{\text{Nevaripine}} - \hat{p}_{\text{Lopinavir}} = 0.217 - 0.083 = 0.134$

$$Z = \frac{0.134 - 0}{0.046} = 2.91$$

برای محاسبه p-value با استفاده از R داریم:

```
> 2*pnorm(2.91,lower.tail = FALSE)
[1] 0.003614288
```

$p - value = P(|Z| > 2.91) = 0.0036 < \alpha = 0.05 \quad \rightarrow \quad \boldsymbol{Reject\ H_0}$

فرض *null* را رد می‌کنیم. یعنی نمی‌توانیم بگوییم که بین درصد *virologic failure* در دو گروه *Nevaripine* و *Lopinavir* اختلافی وجود ندارد. پس اختلاف معناداری بین درصد *virologic failure* در این دو گروه وجود دارد.

---

c) در قسمت b آورده شده

۴-

a) تیم ملی بسکتبال ایران

	Underweight BMI<18.5	Normal Weight BMI 18.5-24.9	Overweight BMI 25.0-29.9	Obese BMI >=30	Total
% in population	2%	39%	36%	23%	100%
Expected #	$3326 \times 0.02 = 66.52 \approx 67$	$3326 \times 0.39 = 1297.14 \approx 1297$	$3326 \times 0.36 = 1197.36 \approx 1197$	$3326 \times 0.23 = 764.98 \approx 765$	3326
Observed #	20	932	1374	1000	3326

از آنجایی که می‌خواهیم دوتا توزیع را باهم مقایسه کنیم می‌توانیم از تست goodness-of-fit استفاده کنیم. شرایط آزمون chi-square را بررسی می‌کنیم:

استقلال:

- نمونه برداری تصادفی انجام شده است✓
- ✓ $3326 < 10\%\ of\ population$
- هر case فقط به یک سلول تعلق دارد(مثلا فردی که اضافه وزن دارد نمی‌تواند هم زمان هم در گروه وزن نرمال هم قرار گرفته باشد) ✓

سایز سمپل:

- هر سناریو(هرسلول) باید حداقل ۵ تا expected case داشته باشد.✓

حال آزمون فرض را طراحی می‌کنیم:

H_0: افراد مشاهده شده از گروه‌های وزن/قد (BMI) مختلف در سمپل Framingham Offspring توزیع BMI یکسانی با جامعه آماری(کل مردم آمریکا در سال ۲۰۰۲) دارند.

H_A: افراد مشاهده شده از گروه‌های وزن/قد (BMI) مختلف در سمپل Framingham Offspring توزیع BMI متفاوتی با جامعه آماری(کل مردم آمریکا در سال ۲۰۰۲) دارند.

$$df = k - 1 = 4 - 1 = 3$$

Significance level را 5% در نظر می‌گیریم (confidence level = 95%)

$$\alpha = 0.05$$

b) از آماره‌ی chi-square استفاده می‌کنیم:

$\chi^2\ statistic$:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E} = \frac{(20-67)^2}{67} + \frac{(932-1297)^2}{1297} + \frac{(1374-1197)^2}{1197} + \frac{(1000-765)^2}{765}$$

$$\chi^2 = 32.97 + 102.72 + 26.17 + 72.19 = 234.05$$

c)

$$Decision\ Rule \begin{cases} if\ p-value < \alpha : reject\ H_0 (the\ 2\ distributions\ aren't\ the\ same) \\ if\ p-value > \alpha : fail\ to\ reject\ H_0 (the\ 2\ distributions\ are\ the\ same) \end{cases}$$

d)در قسمت b حساب شده.

e) حال مقدار p-value را با استفاده از R حساب می‌کنیم:

```
> pchisq(234.05,3,lower.tail = FALSE)
[1] 1.841308e-50
```

$p-value = P(X \geq 234.05|H0) = \ 1.841308e-50 \ \approx 0 < \alpha = 0.05 \ \rightarrow \ \boldsymbol{Reject\ H_0}$

یعنی نمی‌توان گفت که توزیع BMI در سمپل Framingham Offspring با توزیع BMI مردم آمریکا در سال ۲۰۰۲ یکسان می‌باشد.

توزیع سمپل Framingham Offspring و توزیع جامعه آماری متفاوت است.(این جمله را با قطعیت بالایی می‌توان گفت از آنجایی که p-value تقریباً برابر 0 می‌باشد.)

۵- ابتدا شرایط CLT را برای مقایسه‌ی دوتا proportion(درصد پاسخ yes دو گروه را مقایسه می‌کنیم) بررسی می‌کنیم:

استقلال درون‌گروهی: ✓

- Random sampling/assignment ✓
- $70 < 10\%\ all\ car\ drivers\ in\ Iran$, $50 < 10\%\ all\ motorcyclists\ in\ Iran$ ✓

استقلال بین‌گروهی: دو گروه مستقل از هم هستند(در این نمونه هیچ فردی که راننده‌ی ماشین است موتور سوار نیست) ✓

شرط success-failure: ✓ $(\hat{p}_{pool} = \frac{total\ successe}{total\ n} = \frac{23+43}{50+70} = \frac{66}{120} = 0.55\ )$

$n_{motor}\hat{p}_{pool} = 50 \times 0.55 = 27.5\ \geq 10 \quad , \ n_{motor}(1-\hat{p}_{pool}) = \ 50 \times 0.45 = 22.5\ \geq 10$ ✓

$n_{car}\hat{p}_{pool} = 70 \times 0.55 = 38.5\ \geq 10 \quad , \ n_{car}(1-\hat{p}_{pool}) = \ 70 \times 0.45 = 31.5\ \geq 10$ ✓

حال می‌توانیم از آزمون فرض برای مقایسه‌ی دو proportion استفاده کنیم:

H_0: درصد موتورسوارانی که می‌خواهند از بیمه استفاده کنند تا هزینه خسارات خود را بگیرند با درصد رانندگان ماشینی که چنین چیزی را می‌خواهند یکسان است. (درصد پاسخ yes به سوال مطرح شده)

H_A: درصد موتورسوارانی که می‌خواهند از بیمه استفاده کنند تا هزینه خسارات خود را بگیرند با درصد رانندگان ماشینی که چنین چیزی را می‌خواهند متفاوت است. (درصد پاسخ yes به سوال مطرح شده)

$H_0: p_{\text{car}} - \ p_{\text{motor}} = 0$

$H_A: p_{\text{car}} - \ p_{\text{motor}} \neq 0$

$$SE = \sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_{\text{motor}}} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_{\text{car}}}} = \sqrt{\frac{0.55 \times 0.45}{50} + \frac{0.55 \times 0.45}{70}} \approx 0.099$$

$\hat{p}_{\text{car}} - \ \hat{p}_{\text{motor}} \sim N(mean = 0, SE = 0.099)$

$point\ estimate: \hat{p}_{\text{car}} - \ \hat{p}_{\text{motor}} = 0.614 - 0.46 = 0.154$

$$Z = \frac{0.154 - 0}{0.099} = 1.56$$

برای محاسبه p-value با استفاده از R داریم:

```
> 2*pnorm(1.56,lower.tail = FALSE)
[1] 0.1187599
```

$p - value = P(|Z| > 1.56) = \; 0.118 > \alpha = 0.05 \quad \rightarrow \; \boldsymbol{fail\ to\ Reject\ H_0}$

فرض *null* را نمی‌توانیم رد می‌کنیم. یعنی می‌توان گفت *proportion* این دو گروه یکسان است.(با *confidence level = 95%*)

۶-

a) اگر مادر ها واقعا بتوانند بوی بچه‌ی خود را تشخیص بدهند یعنی احتمال تشخیص درست توسط آنها از انتخاب تصادفی($p=\frac{1}{26}=$ 0.038) بهتر است.

$H_0: p = \frac{1}{26} = 0.038$

$H_A: p > 0.038$

b)

	Correct guess	Wrong guess	total
% in population	3.8%	96.2%	100%
Expected #	$0.038 \times 320 = 12.16$	$0.962 \times 320 = 307.84$	320
Observed #	110	210	320

با استفاده از کد R داریم:

```
#b)
#calculating the expected values
correct_guess_observed <- 110
wrong_guess_observed <- 210
correct_guess_expected <- 110*(1/26)
wrong_guess_expected <- 210*(1/26)
#calculating the chi-statistic
chi_square_statistic <- (correct_guess_observed - correct_guess_expected)^2/correct_guess_expected +
  (wrong_guess_observed - wrong_guess_expected)^2/wrong_guess_expected
#calculating degree of freedom
df <- 2-1
##calculating the p-value
p_value <- 0
p_value <- pchisq(chi_square_statistic,df,lower.tail = FALSE)

#outputting the result of the chi_square test
if(p_value < 0.05){
  sprintf("P_value:%s < 0.05  -> mothers don't recognize their child's smell",p_value)
}else{
  sprintf("P_value:%s > 0.05  -> mothers recognize their child's smell",p_value)
}
```

The Result:

```
> #outputting the result of the chi_square test
> if(p_value < 0.05){
+   sprintf("P_value:%s < 0.05  -> mothers don't recognize their child's smell",p_value)
+ }else{
+   sprintf("P_value:%s > 0.05  -> mothers recognize their child's smell",p_value)
+ }
[1] "P_value:0 < 0.05  -> mothers don't recognize their child's smell"
```

همانطور که می‌بینیم مقدار p_value برابر با 0 شده است. پس H0 را رد می‌کنیم و می‌توان گفت که مادرها واقعا می‌توانند بوی بچه‌ی خود را تشخیص دهند.

۷- برای اینکار می‌توان از آزمون استقلال chi-square استفاده کرد. شرایط آزمون chi-square برقرار است(شرط استقلال و شرط sample size). به دو روش اینکار را انجام دادم. یکبار کل آزمون را دستی کد زدم و بار دیگر با تابع chisq.test اینکار را انجام دادم.

روش دستی:

ابتدا سطر و ستون total را به دیتاست caith اضافه می‌کنیم. هر یک از درایه‌های سطر total برابر است با جمع درایه‌های ستون متناظر با آن و هر یک از درایه‌های ستون total برابر است با جمع درایه‌های سطر متناظر با آن.

```
#adding  a "total" column and row to caith
df_caith <- caith
df_caith <- df_caith %>% mutate(total = rowSums(across(where(is.numeric))))
df_caith["total" ,] <- colSums(df_caith[1:4,])
```

سپس جدول Expected Counts را حساب می‌کنیم:

```
#calculating the Expected Counts table as a dataframe
df_caith_expected <-caith
for(i in 1:(length(df_caith$fair)-1)){
  for (j in 1:(length(df_caith)-1)) {
    df_caith_expected[i,j]<- df_caith[i,"total"]*df_caith["total",j]/df_caith["total","total"]
  }
}
print(df_caith_expected)
```

The Result:

```
> print(df_caith_expected)
           fair      red   medium     dark    black
blue   193.9280 38.11918 284.8275 185.3978 15.72749
light  426.7496 83.88342 626.7793 407.9785 34.60924
medium 479.1479 94.18303 703.7383 458.0720 38.85873
dark   355.1745 69.81437 521.6549 339.5517 28.80453
```

سپس آماره‌ی chi-square را حساب می‌کنیم:

```
#calculating the chi-square statistic
chi_square <- 0
for(i in 1:(length(df_caith$fair)-1)){
  for (j in 1:(length(df_caith)-1)) {
    chi_square <- chi_square +
      (df_caith[i,j] - df_caith_expected[i,j])^2/df_caith_expected[i,j]
  }
}
print(chi_square)
```

The Result:

```
> print(chi_square)
[1] 1240.039
```

حال درجه آزادی(df) و P-value را محاسبه کرده و نتیجه‌ی آزمون فرض را به کمک آنها تعیین می‌کنیم:

```
#calculating the degree of freedom
df <- (length(df_caith_expected$fair)-1)*((length(df_caith_expected)-1))

#calculating the p-value
p_value <- 0
p_value <- pchisq(chi_square,df,lower.tail = FALSE)

#outputting the result of the chi_square independence test
if(p_value < 0.05){
  sprintf("P_value:%s < 0.05  ->  Reject H0: eye color and hair color aren't independent",p_value)
}else{
  sprintf("P_value:%s > 0.05  ->  Fail to Reject H0: eye color is independent of hair color",p_value)
}
```

The Result:

```
> #outputting the result of the chi_square independence test
> if(p_value < 0.05){
+   sprintf("P_value:%s < 0.05  ->  Reject H0: eye color and hair color aren't independent",p_value)
+ }else{
+   sprintf("P_value:%s > 0.05  ->  Fail to Reject H0: eye color is independent of hair color",p_value)
+ }
[1] "P_value:4.12399287024362e-258 < 0.05  ->  Reject H0: eye color and hair color aren't independent"
```

همانطور که در خروجی کد مشاهده می‌کنید مقدار p-value خیلی ناچیز(تقریباً 0) می‌باشد و در نتیجه H0 را رد می‌کنیم. یعنی نمی‌توان گفت رنگ چشم و رنگ مو مستقل از همند. و تلویحاً می‌گوییم که رنگ چشم و رنگ مو بهم وابسته‌اند.

روش chisq.test:

```
##second approach
chisq_test <- chisq.test(caith)
chisq_test
chisq_test$expected
```

The Result:

```
	Pearson's Chi-squared test

data:  caith
X-squared = 1240, df = 12, p-value < 2.2e-16

> chisq_test$expected
           fair      red   medium     dark    black
blue   193.9280 38.11918 284.8275 185.3978 15.72749
light  426.7496 83.88342 626.7793 407.9785 34.60924
medium 479.1479 94.18303 703.7383 458.0720 38.85873
dark   355.1745 69.81437 521.6549 339.5517 28.80453
```

۸- اگر در فرض null، احتمال موفقیت متغیر تعریف شده را برابر با 0.5 در نظر بگیریم مانند این است یک سکه سالم داریم و آن را 20 بار پرتاب می‌کنیم و حال میخواهیم ببینیم که چند بار شیر(موفقیت) آماده است(شیر آمدن معادل خراب شدن کامپیوتر در دمای بالاتر از ۱۱۰ درجه می‌باشد)

پس برای آزمون فرض داریم:

$H_0: p = 0.5 \rightarrow$ کامپیوترها در دمای بالاتر از ۱۱۰ درجه آسیب نمی‌بینند و خرابی کامپیوتر در دمای بالاتر از ۱۱۰ درجه اتفاقی بوده است

$H_A: p > 0.5 \rightarrow$ کامپیوترها در دمای بالاتر از ۱۱۰ درجه آسیب می‌بینند

- استفاده از یک سکه سالم و در نظر گرفتن شیر آمدن به عنوان موفقیت(موفقیت = خراب شدن کامپیوتر در دمای بالاتر از ۱۱۰ درجه)
- یک simulation : سکه را ۲۰ بار پرتاب کنید و درصد شیر آمدن را ذخیره کنید($\hat{p}_{sim}$)
- Simulation را ۱۰۰ بار تکرار کنید و درصد شیر آمدن را در هر iteration ذخیره کنید.
- درصد simulation هایی را بدست آورید که در آنها درصد شیر آمدن به اندازه‌ی درصد شیر آمدن مشاهده شده ($\hat{p} = 1$) یا بیشتر از آن باشد.

Proportion مشاهده شده:($\hat{p} = 1$) چون از هر ۲۰ بار پرتاب سکه، هر ۲۰ بار شیر آمده.

```
#H0: p=0.5 , HA: p > 0.5 , p_observed = 1

#simulating 100 times(each simulation 20 coin tosses) and calculating p_value
p_value <- 0
sum_p_value <- 0
set.seed(194830)
for (i in 1:100) {
  p_sim <- sample(c(0,1), replace=TRUE, size=20)
  proportion_sim <- sum(p_sim[TRUE])/20
  if(proportion_sim == 1){
    sum_p_value <- sum_p_value + 1
  }
}
p_value <- sum_p_value/100

#outputting the result of the simulation
if(p_value < 0.05){
  sprintf("P_value:%s < 0.05  ->  Reject H0: computer systems get damaged in
          more than 110 degrees",p_value)
}else{
  sprintf("P_value:%s > 0.05  ->  Fail to Reject H0: computer systems don't get
          damaged in more than 110 degrees",p_value)
}
```

The Result:

```
> #outputting the result of the simulation
> if(p_value < 0.05){
+   sprintf("P_value:%s < 0.05  ->  Reject H0: computer systems get damaged in more than 110 degrees",p_value)
+ }else{
+   sprintf("P_value:%s > 0.05  ->  Fail to Reject H0: computer systems don't get damaged in more than 110 degrees",p_value)
+ }
[1] "P_value:0 < 0.05  ->  Reject H0: computer systems get damaged in more than 110 degrees"
```

همانطور که در نتیجه اجرا آمده، مقدار p-value برابر صفر شده‌است. یعنی در هیچ یک از ۱۰۰ simulation انجام شده، هر ۲۰تا سکه باهم شیر نیامد. این مقدار p-value کوچک یعنی احتمال چنین مشاهده‌ای ($\hat{p} = 1$) با فرض اینکه خرابی سیستم‌ها در دمای بالاتر از ۱۱۰ درجه تصادفی باشد، خیلی ناچیز است. پس H0 را رد می‌کنیم و می‌گوییم خرابی سیستم‌ها در دمای بالاتر از ۱۱۰ درجه نمی‌تواند تصادفی باشد و تلویحاً می‌گوییم سیستم‌های کامپیوتر در دمای بالاتر از ۱۱۰ درجه آسیب می‌بینند.
