Caijun Qin
09/03/2021
HW1

1. a. Response variable: attitude towards gun control
   Explanatory variables: gender, mother's education

   b. Response variable: heart disease
   Explanatory variables: blood pressure, cholesterol level

   c. Response variable: vote for president
   Explanatory variables: race, religion, annual income

2. a. UK political party preference is nominal, because there is no inherent order amongst political parties.

   b. Highest educational degree obtained is ordinal, since there is a prerequisite order of degrees to reach any of the listed degrees.

   c. Patient condition is ordinal as the variable describes how likely the patient will survive, thus the values that can be taken have a ranking.

   d. Hospital location is nominal, since each location is a name with no quantifiable ordering with each other.

   e. Favorite beverage is nominal, because the variable entails choosing one value from a set of options with no information on how (for the rest) each option ranks with each other.

   f. Rating of a movie is ordinal, because ratings are based on a quantitative value – the number of stars – to establish ranking of movies.

3. a. This is initially a multinomial distribution with $n = 100$ observations (or trials) and 4 possible values to choose from per observation. In context, each observation is the correctness of answering a question. Since we care only whether the student chooses the correct answer or not, the distribution can be abstracted to a binomial distribution with the same $n = 100$ observations but $\pi = \frac{1}{4}$ chance of being successful (correct).

   The distribution is denoted by Y~binom(100, 0.25), where *y* is the number of successes.

   b. Yes. The expected value of the distribution is $E[\pi] = np = 100 * 0.25 = 25$. The standard deviation is $\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{100 * \frac{1}{4} * \frac{3}{4}} \approx 4.33$. The actual number of successes, 50, is drastically higher than the expected value. In fact, having 50 correct answers would be $\frac{50-25}{4.33} \approx 5.77$ standard deviations above the mean. With probability, $P(y = 50) = \binom{100}{50}\left(\frac{1}{4}\right)^{50}\left(1-\frac{1}{4}\right)^{100-50} \approx 4.51e-8$, which is almost impossible.

Caijun Qin
09/03/2021
HW1

4. The binomial distribution $Y \sim Bin\left(2, \frac{1}{2}\right)$ has the PMF below.

$P(Y = y) = \binom{2}{y}\left(\frac{1}{2}\right)^{y}\left(1 - \frac{1}{2}\right)^{2-y}$

$\Rightarrow \binom{2}{y}\left(\frac{1}{2}\right)^{y}\left(\frac{1}{2}\right)^{2-y}$

$\Rightarrow \binom{2}{y}\left(\frac{1}{2}\right)^{2}$

$\Rightarrow \frac{\binom{2}{y}}{4}$

The formula $\frac{\binom{2}{y}}{4}$ can be used to compute the probabilities for each value that can be taken by y, namely $y = 0,1,2$. Correspondingly, the probabilities are $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ for the different number of possible values of y.

The mean and standard deviation of the distribution are computed below.

$\mu = E[Y] = n\pi = 2 * \frac{1}{2} = 1$

$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{2 * \frac{1}{2} * \left(1 - \frac{1}{2}\right)} = \frac{\sqrt{2}}{2}$

b. Plugging in the known constants, $n = 2$ and $y = 1$, the likelihood function based on π is derived below.

$L(\pi|y = 1) = \binom{2}{1}\pi^{1}(1-\pi)^{2-1}$

$\Rightarrow L(\pi|y = 1) = 2\pi(1-\pi)$

The MLE $\hat{\pi}$ is then computed below by turning the likelihood function L into an optimization problem i.e. taking the derivative and then set to zero (0). Note that natural log is applied to L before deriving.

$\frac{\partial \ln(L)}{\partial \pi} = \frac{\partial}{\partial \pi}\left[\ln\left(2\pi(1-\pi)\right)\right] = \frac{\partial}{\partial \pi}\left[\ln(2) + \ln(\pi) + \ln(1-\pi)\right] = \frac{1}{\pi} - \frac{1}{1-\pi} = 0$

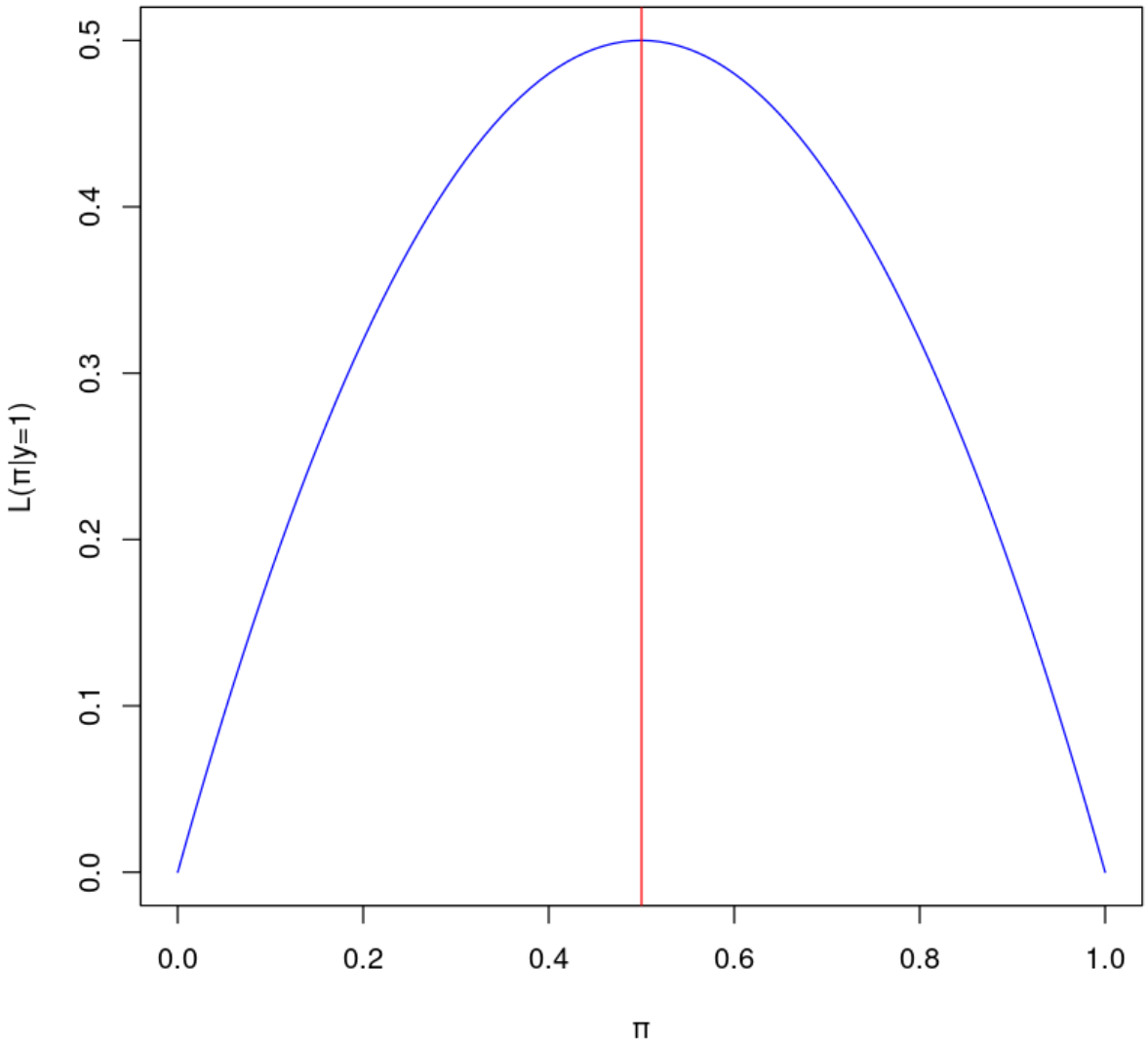
$\Rightarrow \frac{1}{\pi} = \frac{1}{1-\pi} \Rightarrow 1 - \pi = \pi \Rightarrow 1 = 2\pi \Rightarrow \hat{\pi} = \frac{1}{2}$

The R code below can be used to generate $L(\pi|y = 1)$ and the MLE at 0.50.

```
curve(
   expr=2*x*(1-x),
   xlab='π',
   ylab='L(π|y=1)',
   col='blue'
)
```

Caijun Qin
09/03/2021
HW1

```
abline(v=0.5, col='red')
```



As seen in the plot above, the MLE $\hat{\pi} = 0.50$ is that value because the likelihood function L obtains the maximum value over its domain [0, 1]. That is in the definition of MLE. From another perspective, the most likely binomial distribution that assumes initially given data $(n = 2, y = 1)$ has an expected value of 0.50 for successful proportion π.

8. a. $n = 1374$

$\hat{\pi} = \frac{\#\ of\ people\ saying\ yes}{n} = \frac{486}{1374} \approx 0.3537$

Caijun Qin
09/03/2021
HW1

Based on the sample data, approximately 0.3537 of the population would say "yes" to the survey. To construct the 99% confidence interval (CI), the *Wald CI* is used here as an example for the computational steps.

$$CI \equiv \hat{\pi} \pm Z_{1-\frac{\alpha}{2}} * SE \Rightarrow \hat{\pi} \pm Z_{0.995} * \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \Rightarrow 0.3537 \pm 2.807034 * \sqrt{\frac{0.3537(1-0.3537)}{1374}}$$
$$\Rightarrow (0.3175, 0.3899)$$

Using R code below, other types of CI are shown at 99% confidence level.

```
library("binom")
binom.confint(
   x=486,
   n=1374,
   conf.level=0.99,
   method='all'
)
```

	method	x	n	mean	lower	upper
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	agresti-coull	486	1374	0.3537118	0.3212551	0.3875745
2	asymptotic	486	1374	0.3537118	0.3204870	0.3869365
3	bayes	486	1374	0.3538182	0.3208969	0.3872545
4	cloglog	486	1374	0.3537118	0.3205938	0.3869495
5	exact	486	1374	0.3537118	0.3206816	0.3877685
6	logit	486	1374	0.3537118	0.3212347	0.3875974
7	probit	486	1374	0.3537118	0.3210811	0.3874547
8	profile	486	1374	0.3537118	0.3209958	0.3873684
9	lrt	486	1374	0.3537118	0.3210070	0.3873693
10	prop.test	486	1374	0.3537118	0.3285137	0.3797373
11	wilson	486	1374	0.3537118	0.3212625	0.3875671

Regardless of which CI type, a 99% confidence level means that constructing the CI with the same process for repeated samples (in this case, surveys) would yield a CI containing the true proportion π for 99% of the samples. There is a key difference between CI types

though based on guarantee. The Clopper-Pearson CI is "exact" given that at least 99% of repeated samples would generate a CI that contains π. Other CI types e.g. Wald use an approximation of the binomial distribution and does not have such a guarantee.

b. An "exact" test for binomial distribution is used for a two-sided significance test with a confidence level of 99%. Since the prompt is to determine either the majority OR minority of the population would say "yes", a two-sided test is chosen. The null hypothesis is $H_0 = 0.50$. Rejecting H_0 would provide strong evidence that there is a statistically significant majority or minority of people saying "yes".

Use the R code below to compute the p-value.

```
library("stats")
binom.test(
   x=486,
   n=1374,
   p=0.50,
   conf.level=0.99,
   alternative='two.sided'
)
```

```
        Exact binomial test

data:  486 and 1374
number of successes = 486, number of trials = 1374, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
99 percent confidence interval:
 0.3206816 0.3877685
sample estimates:
probability of success
             0.3537118
```

The p-value, less than 2.2e-16, is extremely tiny. Therefore, we reject H_0. The p-value here indicates how likely the difference between the sample proportion and the proposed hypothetical proportion (50%) of people that said "yes" was due to uncontrollable chance. Since that value is tiny, there exists strong evidence that the true proportion is different than 50%. Since the number of people who said "yes" in the sample, 486, was clearly way less than 50% of the sample size, 1374, we can report a minority of the population would say "yes".