Lecture 11: Inference for One Proportion
STAT 630, Fall 2021

Bernoulli random variables and properties of the sample proportion

Let X be a binary random variable following a **Bernoulli distribution** with probability p. We can write this as $X \sim \text{Bern}(p)$. Find the probability mass function, expectation, and variance of X.

Probability mass function:

x	0	1
$P(X = x)$	$1 - p$	p

$$E(X) = \sum_x x \cdot P(X = x) = 0 \cdot (1 - p) + 1 \cdot p = p$$

$$\begin{aligned} Var(X) &= \sum_x (x - E(X))^2 \cdot P(X = x) = (0 - p)^2(1 - p) + (1 - p)^2 p \\ &= p^2(1 - p) + (1 - p)^2 p = p(1 - p) \cdot [p + (1 - p)] = p(1 - p) \end{aligned}$$

Let $X_1, X_2, \cdots, X_n$ be independent random variables (i.e., a random sample) from $\text{Bern}(p)$. For example, the parameter p is the **population proportion** of individuals that support a certain political candidate; $X_i = 1$ if the i-th randomly sampled person votes for the candidate, and $X_i = 0$ otherwise. We can estimate the population proportion with the **sample proportion** defined as:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Find the expectation and variance of the sample proportion:

$$E(\hat{p}) = E\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{np}{n} = p$$

$$Var(\hat{p}) = Var\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) = \frac{np(1 - p)}{n^2} = \frac{p(1 - p)}{n}$$

Central limit theorem for sample proportion

Since the sample proportion $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean of Bernoulli random variables, the central limit theorem gives the following approximation for the sampling distribution of $\hat{p}$:

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

The approximation should only be used if $np \geq 10$ and $n(1-p) \geq 10$. This is often called the "success-failure" condition.

Remark: Let $Y = n\hat{p} = \sum_{i=1}^{n} X_i$, which is the sum of independent Bernoulli random variables. Then the exact distribution Y follows is a binomial distribution with parameters n and p. However, as long as the conditions are satisfied, it is often easier to use the normal approximation.

1-α confidence interval for population proportion p

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

Conditions: $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$, and the sample observations are independent. Generally, the independence condition is satisfied if the data come from a simple random sample.

Derivation:
By the central limit theorem, $Z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim N(0,1)$ approximately. Therefore,

$$P\left(-z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) = 1-\alpha$$

After rearranging terms:

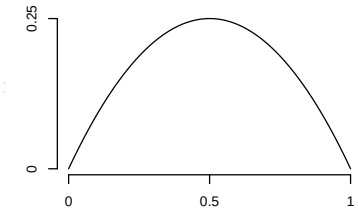
$$P\left(\hat{p} - z_{\alpha/2}\sqrt{p(1-p)/n} < p < \hat{p} + z_{\alpha/2}\sqrt{p(1-p)/n}\right) = 1-\alpha$$

The issue with this interval is the endpoints are in terms of the unknown population proportion parameter p. As an approximation, we replace $\sqrt{p(1-p)/n}$ with $\sqrt{\hat{p}(1-\hat{p})/n}$ when calculating the interval.

Sample size determination
Determine the sample size needed so that the confidence interval will have a margin of error $\pm E$ with confidence level $1-\alpha$.

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \implies E^2 = z^2_{\alpha/2}\frac{\hat{p}(1-\hat{p})}{n} \implies n = \frac{z^2_{\alpha/2}\hat{p}(1-\hat{p})}{E^2}$$

If no data has been collected, $\hat{p} = 0.5$ gives the largest sample size. Why?
Let $f(p) = p(1-p) = -p^2 + p$. Then $f'(p) = -2p + 1$; setting $f'(p) = 0$ gives $p = 1/2$ as the maximum. This can be checked graphically as well since $f(p) = -p^2 + p$ is a parabola facing down.



Ex1: At a survey poll before the elections, candidate A receives the support of 650 voters in a random sample of 1200 voters.

(a) Construct a 95% confidence interval for the population proportion p of voters that support candidate A. Check the conditions for the interval.

(b) Find the sample size needed so that the margin of error will be ± 0.01 with confidence level 0.95.

Solution:
(a) The conditions for the interval are satisfied since $n\hat{p} = 650 \geq 10$ and $n(1-\hat{p}) = 550 \geq 10$, and the voters were random sampled.

The sample proportion is $\hat{p} = 650/1200 = 0.54$ and $n = 1200$.

$$0.54 \pm 1.96\sqrt{\frac{(0.54)(0.46)}{1200}} \implies (0.512, 0.568)$$

We are 95% confident that the population proportion of voters that support candidate A is between 0.512 and 0.568.

(b) Using $\hat{p} = 0.5$ to get the largest sample size:

$$n = \frac{1.96^2(0.5)(1-0.5)}{(0.01)^2} = 9604$$

Hypothesis test for population proportion p

Null and alternative hypotheses:
H_0: $p = p_0$
H_A: $p > p_0$ or $p < p_0$ or $p \neq p_0$

Test statistic: Assuming H_0 is true, $\hat{p} \sim N(p_0, \sqrt{p_0(1-p_0)/n})$. Standardize to get the following z-test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Once the z-test statistic is computed, the p-value can be computed to decide whether to reject or not reject the null hypothesis.

Conditions: $np_0 \geq 10$ and $n(1-p_0) \geq 10$, and independence.

Ex2: A simple random sample of 1,028 US adults in March 2013 found that 56% support nuclear arms reduction. Does this provide convincing evidence that a majority of Americans support nuclear arms reduction?[1]

The conditions for the test are satisfied since $np_0 = n(1-p_0) = 1028(0.5) \geq 10$, and the respondents were randomly sampled.

$H_0 : p = 0.5$
$H_A : p > 0.5$

The sample proportion is $\hat{p} = 0.56$ and $n = 1028$.

Test statistic:

$$z = \frac{0.56 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{1028}}} = 3.85$$

p-value $= P(Z > 3.85) =$ `1 - pnorm(3.85)` ≈ 0

Since the p-value ≈ 0, we reject H_0. The poll provides convincing evidence that a majority of Americans support nuclear arms reduction.

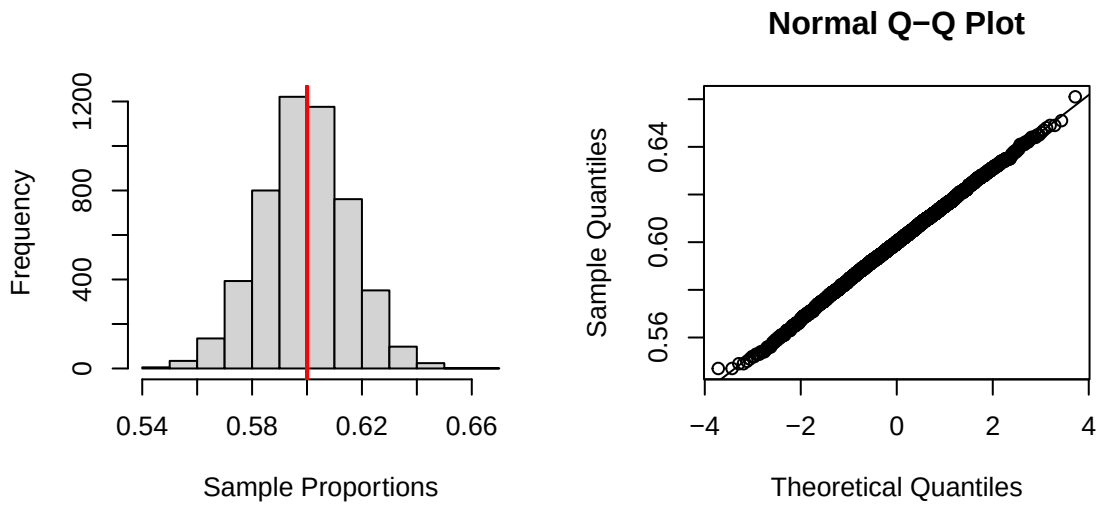
[1] `https://news.gallup.com/poll/161198/favor-russian-nuclear-arms-reductions.aspx`

Simulation Study: Simulating the sampling distribution of the sample proportion when:

- the population size is 1 million
- the sample size is $n = 1000$
- the population proportion is $p = 0.6$

```
set.seed(999)
pop_size <- 10^6
n <- 1000
population <- c(rep(0, 0.4*pop_size), rep(1, 0.6*pop_size))
phats <- c() # initialize vector for sample proportions
for(i in 1:5000) {
  samp <- sample(population, size = n)
  phats[i] <- sum(samp) / n
}
```

```
par(mfrow = c(1, 2), cex=0.8)
hist(phats, xlab = "Sample Proportions", main = "")
abline(v = 0.6, col = "red", lwd = 2)
qqnorm(phats)
qqline(phats)
```




```
sd(phats)
```

```
## [1] 0.01578796
```

```
sqrt(0.6 * 0.4 / 1000)
```

```
## [1] 0.01549193
```