STAT011 Statistical Methods I

Lecture 18 Inference for a Proportion

Lu Chen
Swarthmore College
4/2/2019

Review - Inferences for population means

Inference for	μ **(σ known)**	μ **(σ unknown)**	$\mu_1 - \mu_2$ **($\sigma_1 \neq \sigma_2$)**	$\mu_1 - \mu_2$ **($\sigma_1 = \sigma_2$)**
Name	One-sample z procedures	One-sample t procedures (Paired two-sample t procedures)	Two-sample t procedures	Pooled two-sample t procedures
Based on	$N(0, 1)$	$t(n-1)$	$t(k)$	$t(n_1 + n_2 - 2)$
Estimate	$\bar{x}$	$\bar{x}$	$\bar{x}_1 - \bar{x}_2$	$\bar{x}_1 - \bar{x}_2$
Level C C.I.	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$	$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$\bar{x}_1 - \bar{x}_2 \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

k is computed by Welch-Satterthwaite formula or the smaller of $n_1 - 1$ and $n_2 - 1$.

Review - Inference for population means

Inference for	μ **(σ known)**	μ **(σ unknown)**	$\mu_1 - \mu_2$ **($\sigma_1 \neq \sigma_2$)**	$\mu_1 - \mu_2$ **($\sigma_1 = \sigma_2$)**
Name	One-sample z procedures	One-sample t procedures (Paired two-sample t procedures)	Two-sample t procedures	Pooled two-sample t procedures
H_0	$\mu = \mu_0$	$\mu = \mu_0$	$\mu_1 = \mu_2$	$\mu_1 = \mu_2$
Test statistic	$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \overset{approx.}{\sim} N(0,1)$	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \overset{approx.}{\sim} t(n-1)$	$t = \frac{\bar{x}_1-\bar{x}_2-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \overset{approx.}{\sim} t(k)$	$t = \frac{\bar{x}_1-\bar{x}_2-0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \overset{approx.}{\sim} t(n_1+n_2-2)$

k is computed by Welch-Satterthwaite formula or the smaller of $n_1 - 1$ and $n_2 - 1$.

Outline

- Motivation example: an analysis of many statistical analyses
- Data
- Bernoulli distribution
- Sampling distribution of a sample proportion
- Inference for a population proportion
  - p is unknown…
  - Large sample C.I. for a population proportion
  - Large sample z test for a population proportion
- Examples

Motivation example (Link)

Multiple Researchers Examining the Same Data Find Very Different Results

A new study demonstrates how the choice of statistical techniques when examining data plays a large role in scientific outcomes.

Peter Simons • November 6, 2018

If quantitative psychological science delivers objective facts, then it might be assumed that several different quantitative researchers examining the same data set would come to the same results. Unfortunately, it appears that this is not the case. A new study finds that the various choices made by researchers in the statistical analysis can lead to different results, even when analyzing the same data set.

Motivation example (Link)

In the study, twenty-nine teams, made up of a total of 61 international researchers, were given the same data and each group was asked to conduct an analysis of the data. The question was relatively straightforward—are soccer referees more likely to give red cards to players with darker skin than to those with lighter skin?

The answer, on the other hand, was not as straightforward. Twenty teams (69%) found a significant effect (referees were more likely to give red cards to darker-skinned players), while nine teams (31%) found that there was not a significant effect (referees did not appear to discriminate according to skin tone). Even amongst the researchers that found a positive result, the effect ranged from very slight to very large.

So how did these researchers—looking at the exact same data—arrive at such different results? The answer lies in the choice of statistical analysis and the covariates examined by the researchers.

Nature Comment and the Paper

Data - Exploratory data analysis

29 teams conducted statistical analyses on the same data set; 20 teams found a significant effect; 9 teams found no significant effect.

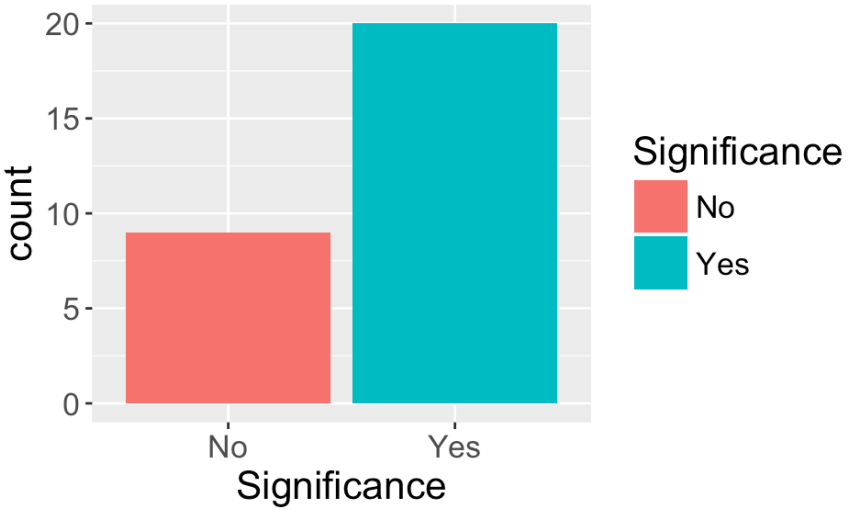
```
head(RedCards)
```

```
##   Team Significance
## 1    1          Yes
## 2    2          Yes
## 3    3          Yes
## 4    4           No
## 5    5          Yes
## 6    6           No
```

```
table(RedCards$Significance)
```

```
##
##  No Yes
##   9  20
```

```
library(ggplot2)
ggplot(RedCards,
       aes(Significance, fill=Significance))+
  geom_bar()+
  theme(text=element_text(size=16)) # optional
```



Data - Question of interest

Question of interest

- What is the 95% confidence interval for the proportion of teams who found a significant effect?
- Is the proportion of teams who found a significant effect different from 0.5, which suggests that the results from the 29 teams are no better than random guess?

Notations

- **Sample size** $n = 29$
- **Population proportion** of teams who found a significant effect: p, unknown
- **Sample proportion** of teams who found a significant effect: $\hat{p} = \frac{20}{29} = 0.690$

Bernoulli distribution

	Mean	Standard deviation	Proportion
Population Parameter	μ	σ	p
Sample Statistic	$\bar{x}$	s	$\hat{p}$

- A dummy variable X with values 0 and 1 follows a **Bernoulli distribution**

$$X \sim Bernoulli(p)$$

- The proportion of $X = 1$ (usually called success) is p.
- The proportion of $X = 0$ (usually called failure) is $1 - p$.
- The mean and SD of X are $\mu_X = p$ and $\sigma_X = \sqrt{p(1-p)}$.
- The sample proportion $\hat{p}$ can be considered as a sample mean and thus a special case of $\bar{x}$.

Sampling distribution of a proportion

Central Limit Theorem

- Population distribution is Normal, $X \sim N(\mu, \sigma)$,

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- Population distribution is not Normal, $\mu_X = \mu, \sigma_X = \sigma$,

$$\bar{x} \overset{approx.}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

  - Population distribution is Bernoulli, $X \sim Bernoulli(p)$, $\mu_X = p$, $\sigma_X = \sqrt{p(1-p)}$,

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

Sampling distribution of a proportion

Normal approximation for proportions

Draw an SRS of size n from a large population having population proportion p of successes. Let $\hat{p}$ be the sample proportion of successes. When n is large, the sampling distribution of $\hat{p}$ is approximately Normal with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$:

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

As a rule of thumb, we will use this approximation for values of n and p that satisfy $np \geq 10$ and $n(1-p) \geq 10$.

Inference for a proportion

Inference for a population mean:

- $\bar{x} \overset{approx.}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$; We make inferences about the unknown μ using $\bar{x}$.
- **Level C CI**: $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$; **Level α test**: $H_0 : \mu = \mu_0, z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$
- When σ is unknown, replace it with sample SD s.

Inference for a population proportion:

- $\hat{p} \overset{approx.}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$; We make inferences about the **unknown** p using $\hat{p}$.
- **Level C CI**: $\hat{p} \pm z^* \sqrt{\frac{p(1-p)}{n}}$; **Level α test**: $H_0 : p = p_0, z = \frac{\hat{p}-p_0}{\sqrt{\frac{p(1-p)}{n}}}$
- Since p is **unknown**, there is NO WAY to calculating $\sqrt{\frac{p(1-p)}{n}}$.

Inference for a proportion

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

- For **confidence interval**, we plug in $\hat{p}$ to calculate $\text{SD}_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ as

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- For **significance test**, since $H_0 : p = p_0$, we plug in p_0 to calculate $\text{SD}_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ as

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

Large sample C.I. for a population proportion

Choose an SRS of size n from a large population with an unknown proportion p of successes. The **sample proportion** is $\hat{p} = \frac{X}{n}$, where X is the number of successes. The **standard error of $\hat{p}$** is

$$\mathrm{SE}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

and the **margin of error** for confidence level C is $m = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where the critical value z^* is the value for the standard Normal density curve with area C between $-z^*$ and z^*. An **approximate level C confidence interval for p** is

$$\hat{p} \pm m = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

for $X \geq 10$ and $n - X \geq 10$.

Large sample z test for a population p

Draw an SRS of size n from a large population with an unknown proportion p of successes. To test the hypothesis $H_0 : p = p_0$, compute the **z statistic**

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{approx.}{\sim} N(0, 1)$$

In terms of a standard Normal random variable Z, the approximate P-value for a test of H_0 against

$$H_a : p > p_0 \text{ is } P(Z \geq z)$$
$$H_a : p < p_0 \text{ is } P(Z \leq z)$$
$$H_a : p \neq p_0 \text{ is } 2P(Z \geq |z|)$$

We recommend the large-sample z significance test as long as $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.

Inference for a population proportion

Note:

- $\hat{p} \overset{approx.}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ if $np \geq 10$ and $n(1-p) \geq 10$.
- When calculating confidence interval for p
  - $\text{SE}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
  - The condition is number of successes $X \geq 10$ and number of failures $n - X \geq 10$.
- When calculating the z test statistic, since we **assume $H_0 : p = p_0$ is true**,
  - $\text{SE}_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$.
  - The condition is, under H_0, number of successes $np_0 \geq 10$ and number of failures $n(1-p_0) \geq 10$.

Example 1: an analysis of 29 statistical analyses

$$n = 29, X = 20, \hat{p} = \frac{X}{n} = \frac{20}{29} = 0.690$$

- **95% confidence interval**

  $\hat{p} \pm m = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.690 \pm 1.96\sqrt{\frac{0.690(1-0.690)}{29}} = 0.690 \pm 0.168$

  $z^* =$ `qnorm(0.975)`

  We are 95% confident that the interval [0.522, 0.858] will contain the true proportion of teams who found a significant effect.

- **Level 0.05 test** $H_0 : p = 0.5$ vs. $H_a : p \neq 0.5$

  $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.690-0.5}{\sqrt{0.5(1-0.5)/29}} = \frac{0.190}{0.093} = 2.05$

  $z > 1.96$ or $P = 2P(Z \geq 2.05) = 0.040 < 0.05$; `2*(1-pnorm(2.05))`

  The test is marginally significant at level 0.05 so we reject H_0. The true proportion of teams who found a significant effect is significantly different from 0.5.

Example 1: an analysis of 29 statistical analyses

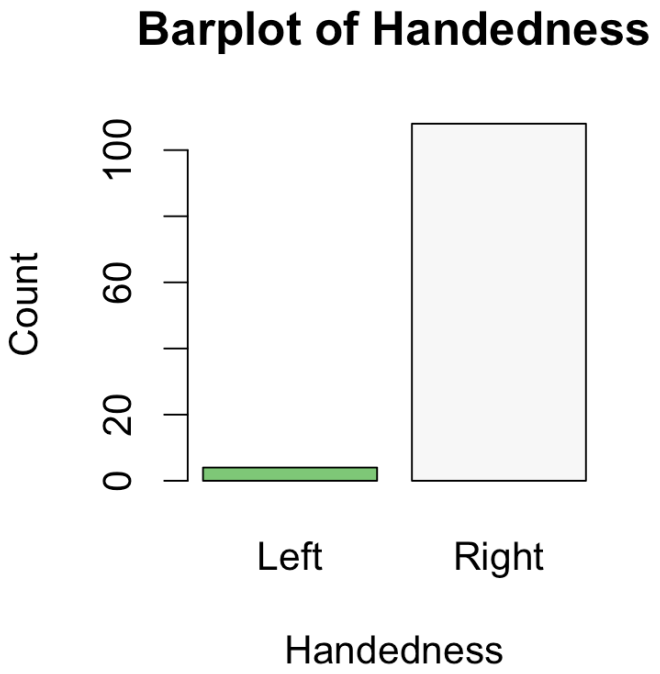
```r
prop.test(x = 20, n = 29, p = 0.5, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  20 out of 29, null probability 0.5
## X-squared = 4.1724, df = 1, p-value = 0.04109
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5076987 0.8272438
## sample estimates:
##         p
## 0.6896552
```

- 95% CI: $[0.508, 0.827]$. The method used to calculate this CI in R is different from ours and thus the interval is slightly different.
- 0.05 test: "X-squared" is the square of our z statistic. P-value is slightly different because of rounding errors.

Example 2 - Left-handedness

2019 Class	Left	Right	Total
Count	4	108	112
Proportion	0.036	0.964	1



- Is the STAT 11 proportion of left-handedness different from the US proportion 0.118?
- $n = 112, X = 4, \hat{p} = 0.036$
- 95% CI: $0.036 \pm 1.96\sqrt{0.036(1-0.036)/112}$ $= 0.036 \pm 0.035$. We are 95% confident that the interval $[0.001, 0.071]$ will contain the true proportion of left-handedness in STAT 11.
- 0.05 test: $H_0 : p = 0.118$ vs. $H_a : p \neq 0.118$. $z = \frac{0.036-0.118}{\sqrt{0.118(1-0.118)/112}} = -2.69 < -1.96$ or $P = 0.007 < 0.05$. the true proportion of left-handedness in STAT 11 is significantly different from the US proportion of left-handedness.

Example 2 - Left-handedness

```
prop.test(x = 4, n = 112, p = 0.118, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  4 out of 112, null probability 0.118
## X-squared = 7.2865, df = 1, p-value = 0.006948
## alternative hypothesis: true p is not equal to 0.118
## 95 percent confidence interval:
##  0.01397461 0.08824664
## sample estimates:
##          p
## 0.03571429
```

Example - Coin toss

Suppose a coin is tossed for 40 times and there are 26 heads. Use 95% confidence interval and significance test to see whether this is a fair coin. Do the two inference methods give you the same conclusion?

- $X = 26, \hat{p} = 26/40 = 0.65, n = 40$
- **95% confidence interval**

  $0.65 \pm 1.96\sqrt{\frac{0.65\times0.35}{40}} = 0.65 \pm 0.148$

  The interval $[0.502, 0.798]$ does not contain 0.5. This may not be a fair coin.
- **Level 0.05 test** $H_0 : p = 0.5, H_a : p \neq 0.5$

  $z = \frac{0.65-0.5}{\sqrt{0.5\times0.5/40}} = 1.897$

  $-1.96 < z < 1.96$ and $P = 0.058 > 0.05$. This may be a fair coin.
- Note: because the computations for the $SE_{\hat{p}}$ are different in the two inference methods, the results from the interval and the test are usually the same but sometimes slightly different.

Example - Coin toss

```
prop.test(x = 26, n = 40, p = 0.5, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  26 out of 40, null probability 0.5
## X-squared = 3.6, df = 1, p-value = 0.05778
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4950588 0.7786547
## sample estimates:
##    p
## 0.65
```

- The methods used by R for computing the CI and conducting the test resulted in the same conclusion.

Sampling distributions

Sample mean	$\bar{x}$, Mean	$\hat{p}$, Proportion of sucesses
Computation	$\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n}$	$\hat{p} = \frac{X}{n}$ X is the number of sucesses
Mean	μ	p
SD	$\frac{\sigma}{\sqrt{n}}$	$\sqrt{\frac{p(1-p)}{n}}$
Distribution	$\bar{x} \overset{approx.}{\sim} N\left(\mu, \frac{\sigma}{n}\right)$	$\hat{p} \overset{approx.}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$
SE	$\frac{s}{\sqrt{n}}$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ or $\sqrt{\frac{p_0(1-p_0)}{n}}$

One-sample procedures

Inference for	μ σ is known	μ σ is unknown	p p is unknown
Name	One-sample z	One-sample t	Inference for a proportion
Based on	$N(0, 1)$	$t(k)$	$N(0, 1)$
Estimate	$\bar{x}$	$\bar{x}$	$\hat{p}$
Level C C.I.	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$	$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Critical values	z^* `qnorm(1-(1-C)/2)`	t^* `qt(1-(1-C)/2,df=n-1)`	z^* `qnorm(1-(1-C)/2)`
H_0	$\mu = \mu_0$	$\mu = \mu_0$	$p = p_0$
Test statistic	$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \overset{approx.}{\sim} N(0, 1)$	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \overset{approx.}{\sim} t(n-1)$	$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{approx.}{\sim} N(0, 1)$

Summary

- Motivation example: an analysis of many statistical analyses
- Data
- Bernoulli distribution $X \sim Bernoulli(p)$
- Sampling distribution of a sample proportion $\hat{p} \overset{approx.}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$
- Inference for a population proportion
  - p is unknown…
  - Large sample C.I. for a population proportion $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - Large sample z test for a population proportion $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{approx.}{\sim} N(0, 1)$
- Examples