Chapter 5: Foundations for inference

Wen-Han Hwang
(Slides primarily developed by Mine Çetinkaya-Rundel from OpenIntro.)

Institute of Statistics
National Tsing Hua University
Taiwan

OpenIntro

Outline

Point estimates and sampling variability

Point estimates and error

- We are often interested in population parameters, *e.g.*, p, μ, σ.
- Complete populations are difficult to collect data on, so we use sample statistics as point estimates for the unknown population parameters of interest, e.g. $\hat{p}, \bar{x}, s$.
- Error in the estimate, *e.g.*, $\hat{p} - p, \bar{x} - \mu, s - \sigma$.
- Bias is systematic tendency to over- or under-estimate the true population parameter, *e.g.*, $E(\hat{p} - p), E(\bar{x} - \mu), E(s - \sigma)$.
- Sampling error describes how much an estimate will tend to vary from one sample to the next, *e.g.*, $\sqrt{\text{Var}(\hat{p})}, \sqrt{\text{Var}(\bar{x})}, \sqrt{\text{Var}(s)}$.
- Much of statistics is focused on understanding and quantifying sampling error, and sample size is helpful for quantifying this error.

New Project 2024 03 25t111704.946

臺灣調查網記者徐以琳 / 綜合報導

近日美國蓋洛普民調(Gallup Poll)公布了民眾對總統拜登的看法調查，結果發現，雖然距離總統大選還有一段時間，但由於拜登目前支持率未過五成，且美人普遍對國家發展現狀不滿意，因此可見得其選情危急，能否成功連任仍是未知數。

最近美國因為2024總統大選而陷入一波熱潮，其中並以川普拜登之戰最受選民注目，然根據美國蓋洛普民調公布的民調顯示，如今大多數美國民眾都認為拜登無能力再連任，顯示其如今聲望未回漲，美國民眾對政府的施政能力也相當不滿意。

調查顯示，當前在美國，僅有40%的民眾對拜登的施政表現表示支持，相比於川普、歐巴馬2020年、2012年同期的47%、46%民調還分別低了7、6個百分點，且上述兩人無一例外地都未成功連任，顯示此次支持率同樣未過五成的拜登連任路危殆，選情岌岌可危。

此外，當問及對拜登處理國際事務的看法時，也僅有47%的美國民眾對拜登處理中東局勢的反應感到滿意，加上另外僅37%的美人表示對其處理國內經濟的表現很滿意，42%的該國民眾並對其在能源政策方面的施政表現感到滿意，顯示拜登如今確實深陷低迷民意，勝選機率被迫再度下修。

本次美國民眾對總統拜登的看法調查，是美國蓋洛普民調(Gallup Poll)於3月22日公布、~~3月1日至20日~~執行的民意調查，針對美國18歲以上的成年民眾進行線上調查，共完成1016份有效樣本，在95%的信賴水準下，抽樣誤差為正負4.0個百分點。

照片來源：拜登臉書

Young, Underemployed and Optimistic

Coming of Age, Slowly, in a Tough Economy

Young adults hit hard by the recession. A plurality of the public (41%) believes young adults, rather than middle-aged or older adults, are having the toughest time in today's economy. An analysis of government economic data suggests that this perception is correct. The recent indicators on the nation's labor market show a decline in the

Tough economic times altering young adults' daily lives, long-term plans. While negative trends in the labor market have been felt most acutely by the youngest workers, many adults in their late 20s and early 30s have also felt the impact of the weak economy. Among all 18- to 34-year-olds, fully half (49%) say they have taken a job they didn't want just to pay the bills, with 24% saying they have taken an unpaid job to gain work experience. And more than one-third (35%) say that, as a result of the poor economy, they have gone back to school. Their personal lives have also been affected: 31% have postponed either getting married or having a baby (22% say they have postponed having a baby and 20% have put off getting married). One-in-four (24%) say they have moved back in with their parents after living on their own.

http://pewresearch.org/pubs/2191/young-adults-workers-labor-market-pay-careers-advancement-recession

Margin of error

The general public survey is based on telephone interviews conducted Dec. 6-19, 2011, with a nationally representative sample of 2,048 adults ages 18 and older living in the continental United States, including an oversample of 346 adults ages 18 to 34. A total of 769 interviews were completed with respondents contacted by landline telephone and 1,279 with those contacted on their cellular phone. Data are weighted to produce a final sample that is representative of the general population of adults in the continental United States. Survey interviews were conducted under the direction of Princeton Survey Research Associates International, in English and Spanish. Margin of sampling error is plus or minus 2.9 percentage points for results based on the total sample and 4.4 percentage points for adults ages 18-34 at the 95% confidence level.

- 41% $\pm$ 2.9%: We are 95% confident that 38.1% to 43.9% of the public believe young adults, rather than middle-aged or older adults, are having the toughest time in today's economy.
- 49% $\pm$ 4.4%: We are 95% confident that 44.6% to 53.4% of 18-34 years olds have taken a job they didn't want just to pay the bills.

Application exercise

Suppose the proportion of American adults who support the expansion of solar energy is $p = 0.88$, which is our parameter of interest. Is a randomly selected American adult more or less likely to support the expansion of solar energy?

More likely.

Application exercise

Suppose that you don't have access to the population of all American adults, which is a quite likely scenario. To estimate the proportion of American adults who support solar power expansion, you might sample from the population and use your sample proportion as the best guess for the unknown population proportion.

- Sample 1000 American adults from the population, and record whether they support or not solar power expansion.
- Find the sample proportion.
- Plot the distribution of the sample proportions obtained by members of the class.

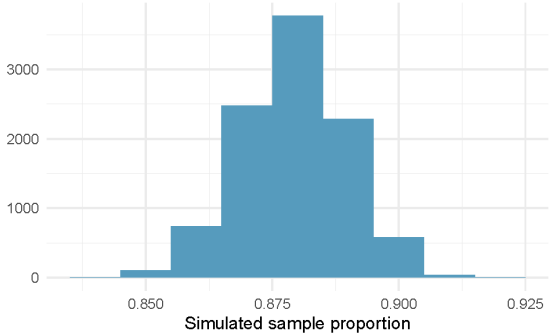
```r
# 1. Create a set of 250 million entries, where 88% of # them are "support" and 12% are "not".
pop_size <- 250000000
entries <- c(rep("support", 0.88 * pop_size),  rep("not", 0.12 * pop_size))

# 2. Initialize a vector to store p-hat values.
p_hat_values <- numeric(500)

# 3. Replicate the sampling process 1000 times.
for (i in 1:1000) {
  sampled_entries <- sample(entries, size = 1000,  replace = F)
  p_hat_values[i] <- sum(sampled_entries == "support") / 1000}

# 4. Plot the histogram of p-hat values.
hist(p_hat_values, main = "Sampling Dist of p-hat",  xlab = "p-hat", breaks = 20)
```

Repeat 1000 times

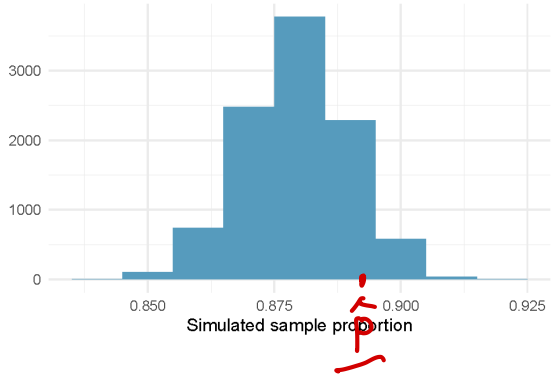
1000's $\hat{p}$

Sampling distribution

When we replicate the sampling process multiple times, each time computing a new proportion ($\hat{p}$) of "support" entries, we generate a collection of $\hat{p}$ values.
This collection forms what is known in statistics as the sampling distribution of $\hat{p}$.



What is the shape and center of this distribution? Based on this distribution, what do you think is the true population proportion?



The distribution is unimodal and roughly symmetric. A reasonable guess for the true population proportion is the center of this distribution, approximately 0.88.

Sampling distributions are never observed

- In real-world applications, we never actually observe the sampling distribution, yet it is useful to always think of a point estimate as coming from such a hypothetical distribution.
- Understanding the sampling distribution will help us characterize and make sense of the point estimates that we do observe.

Central Limit Theorem

$np \geq 5$, $n(1-p) \geq 5$

Central limit theorem

$X \sim Bino(n, p)$

$\hat{p} = \frac{X}{n}$, $E(\hat{p}) = p$

Sample proportions will be nearly normally distributed with mean equal to the population proportion, p, and standard error equal to $\sqrt{\frac{p\,(1-p)}{n}}$.

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

$Var(\hat{p}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$

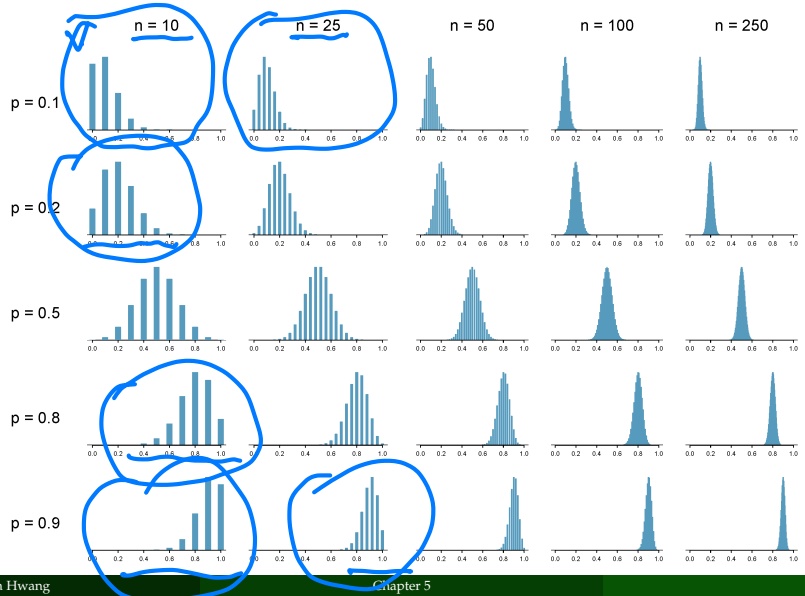
- It wasn't a coincidence that the sampling distribution we saw earlier was symmetric, and centered at the true population proportion.
- We won't go through a detailed proof of why $SE = \sqrt{\frac{p\,(1-p)}{n}}$, but note that as n increases SE decreases.
  - As n increases samples will yield more consistent $\hat{p}$s, i.e. variability among $\hat{p}$s will be lower.

CLT - conditions

Certain conditions must be met for the CLT to apply:

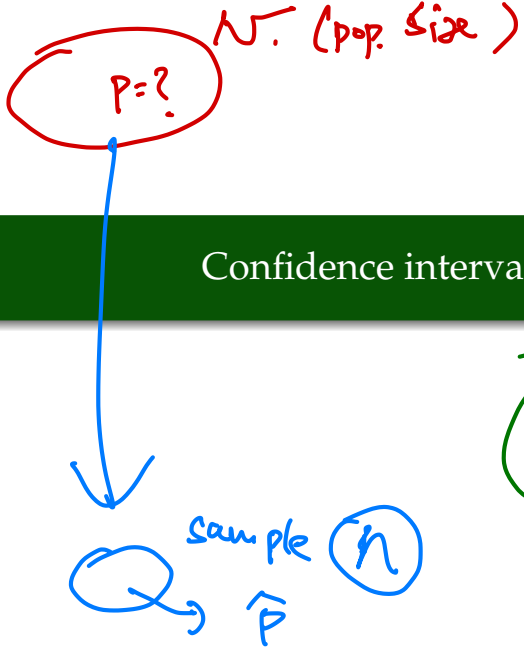
1. Independence: Sampled observations must be independent.
   This is crucial for ensuring that the sampling distribution reflects the population.
   Independence is more likely if:
   - Random sampling or assignment is employed.
   - When sampling without replacement, the sample size (n) is less than 10% of the population size.
2. Sample Size: The sample must be large enough to expect at least 5 successes ($n \times p$) and 5 failures ($n \times (1-p)$) in the sample.
   - Since p (the true population proportion) is often unknown, we use observed successes and failures in the sample, substituting $\hat{p}$ for p, to assess this condition.

What happens when np and/or $n(1-p) < 5$?



Extending the framework for other statistics

- The strategy of using a sample statistic to estimate a parameter is quite common, and it's a strategy that we can apply to other statistics besides a proportion.
  - Take a random sample of students at a college and ask them how many extracurricular activities they are involved in to estimate the average number of extra curricular activities all students in this college are interested in.
- The principles and general ideas are from this chapter apply to other parameters as well, even if the details change a little.



Confidence intervals for a proportion

Confidence intervals

- A plausible range of values for the population parameter is called a confidence interval.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

- If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Facebook's categorization of user interests

- Commercial websites utilize user behavior data for targeted content and ads.
- Pew Research surveyed 850 American Facebook users to evaluate the accuracy of Facebook's interest categorization.
- Question: Do the categories Facebook assigns align with users' actual interests?
- Finding: 67% of respondents affirmed the accuracy of their interest categories.
- Objective: Estimate the true proportion of American Facebook users satisfied with their interest categorization.

https://www.pewinternet.org/2019/01/16/facebook-algorithms-and-personal-data/

Facebook's categorization of user interests

$$\hat{p} = 0.67 \qquad n = 850$$

The approximate 95% confidence interval is defined as

$$point\ estimate \pm 1.96 \times SE$$

$$\hat{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.67 \times 0.33}{850}} \approx 0.016$$

$$\begin{aligned} \hat{p} \pm 1.96 \times SE &= 0.67 \pm 1.96 \times 0.016 \\ &= (0.67 - 0.03, 0.67 + 0.03) \\ &= (0.64, 0.70) \end{aligned}$$

CLT

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

σ^2

$$P\left(\frac{|\hat{p}-p|}{\sigma} \leq 2\right) \approx 95\%$$

$$P\left(p \in \hat{p} \pm 2 \cdot \hat{\sigma}\right) \approx 95\%$$

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

$\hat{\sigma}$: Sampling error

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

(a) 64% to 70% of American Facebook users in this sample think Facebook categorizes their interests accurately.

$p \in (0.64, 0.70)$

(b) 64% to 70% of all American Facebook users think Facebook categorizes their interests accurately

(c) there is a 64% to 70% chance that a randomly chosen American Facebook user's interests are categorized accurately.

(d) there is a 64% to 70% chance that 95% of American Facebook users' interests are categorized accurately.

5% CI

$$\hat{p} \pm 2 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

confidence level

99.7%

$$\hat{p} \pm 3 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

99% CI : $\hat{p} \pm (2.58) \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

What does 95% confident mean?

- Suppose we took many samples and built a confidence interval from each sample using the equation $point\ estimate \pm 1.96 \times SE$.
- Then about 95% of those intervals would contain the true population proportion (p).

Suppose we took many samples and built a 95% confidence interval from each. Then about 95% of those intervals would contain the parameter, p. Figure shows the process of creating 25 intervals from 25 samples from the simulation in Section 5.1.2, where 24 of the resulting confidence intervals contain the simulation's population proportion of $p = 0.88$, and one interval does not.

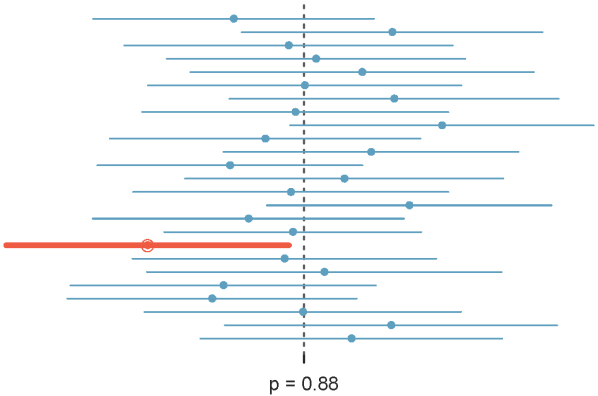
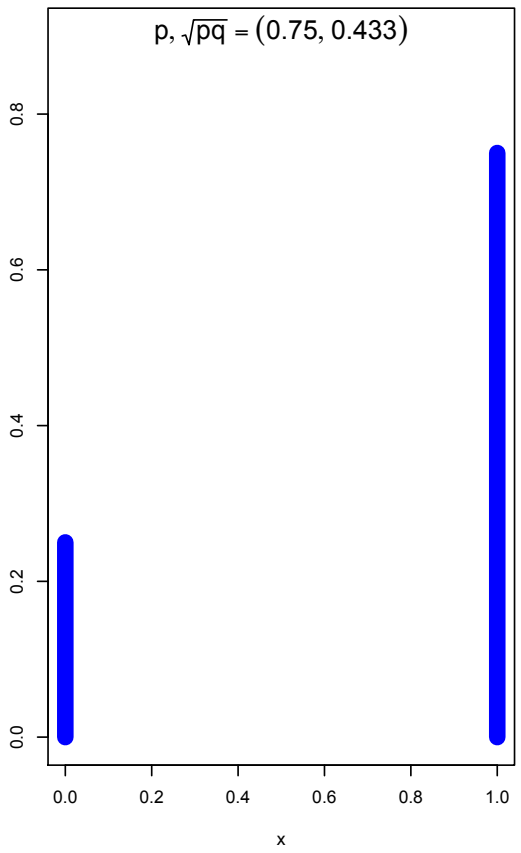
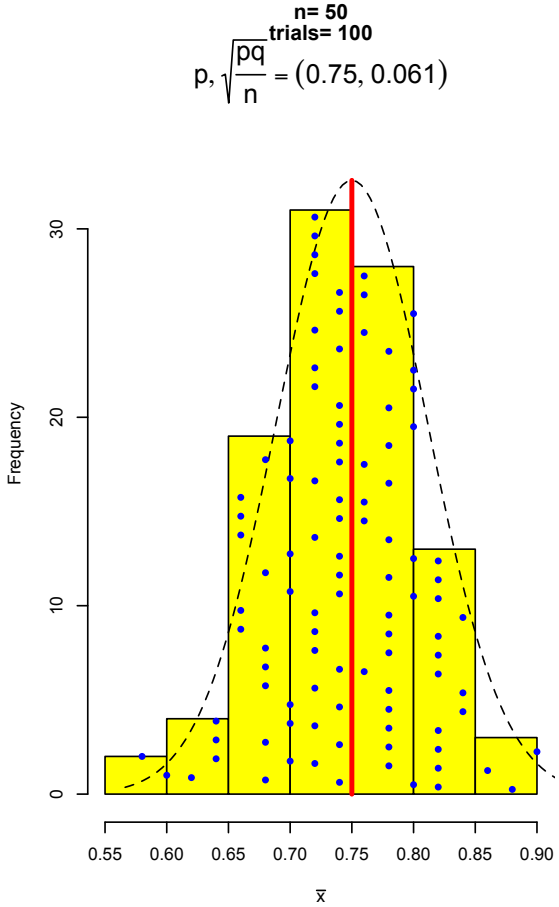
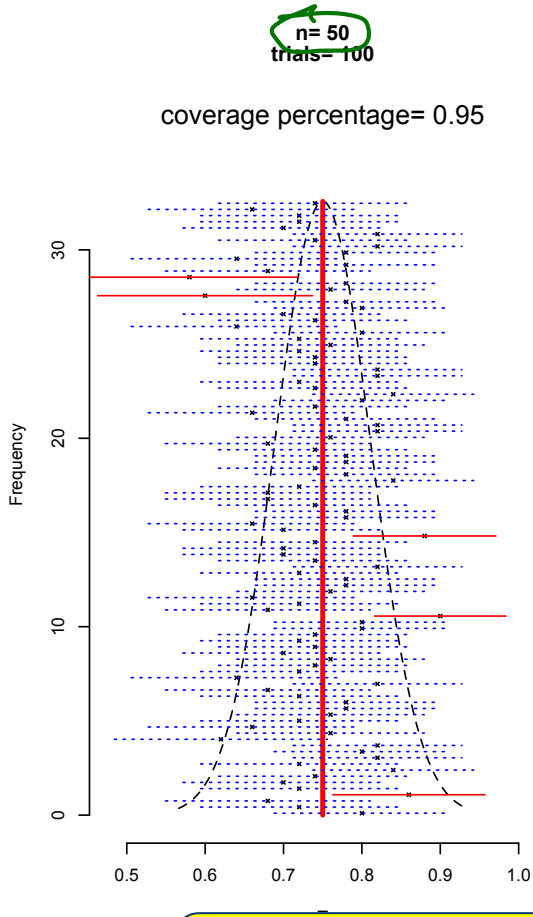


Figure: Twenty-five point estimates and confidence intervals from the simulations in Section 5.1.2. These intervals are shown relative to the population proportion $p = 0.88$. Only 1 of these 25 intervals did not capture the population proportion, and this interval has been bolded.

p 的95%信賴區間模擬驗證
n=50, 母體的p=75%

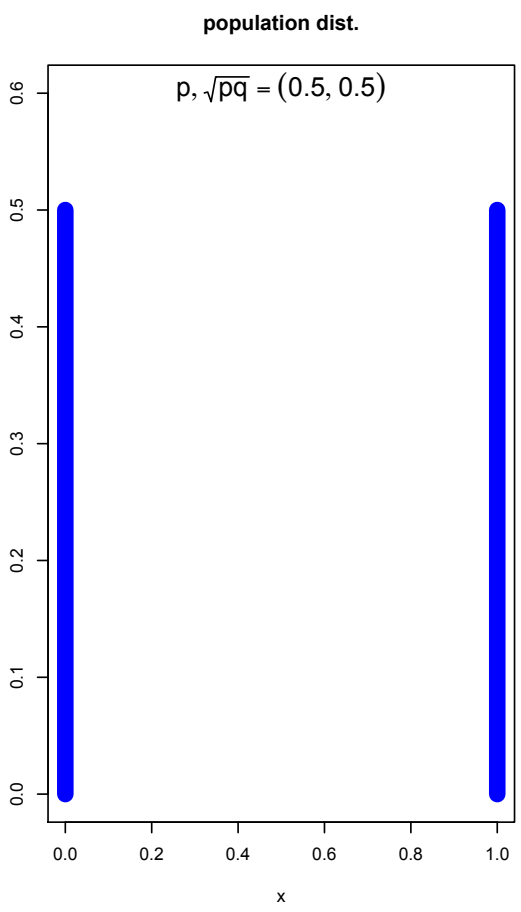


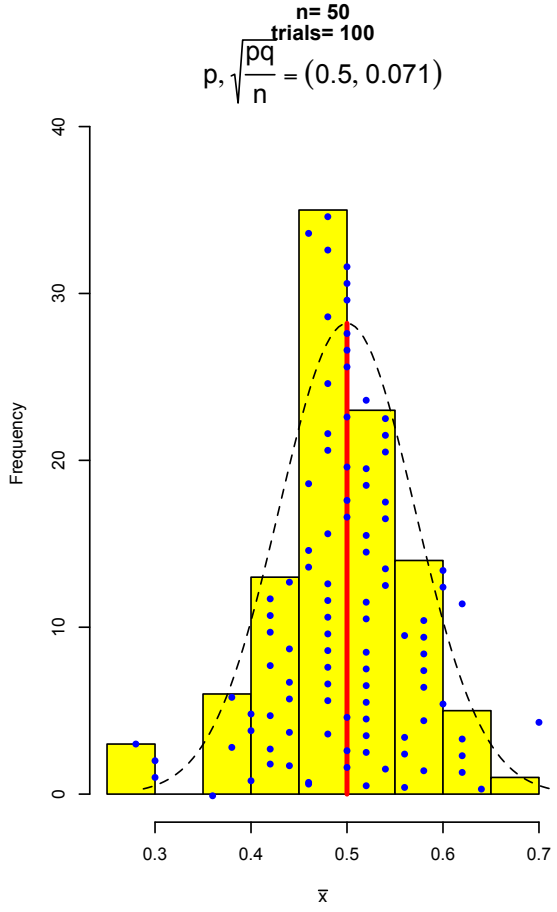


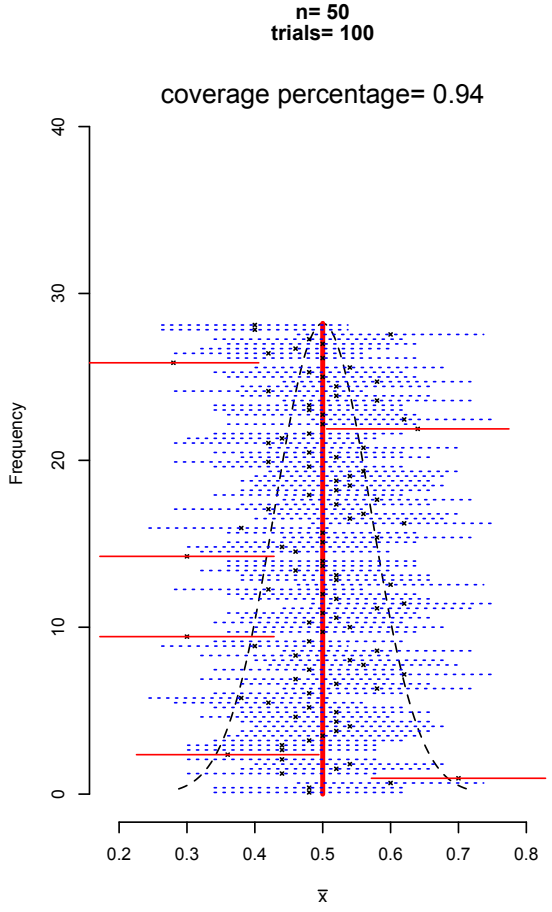


區間涵蓋率

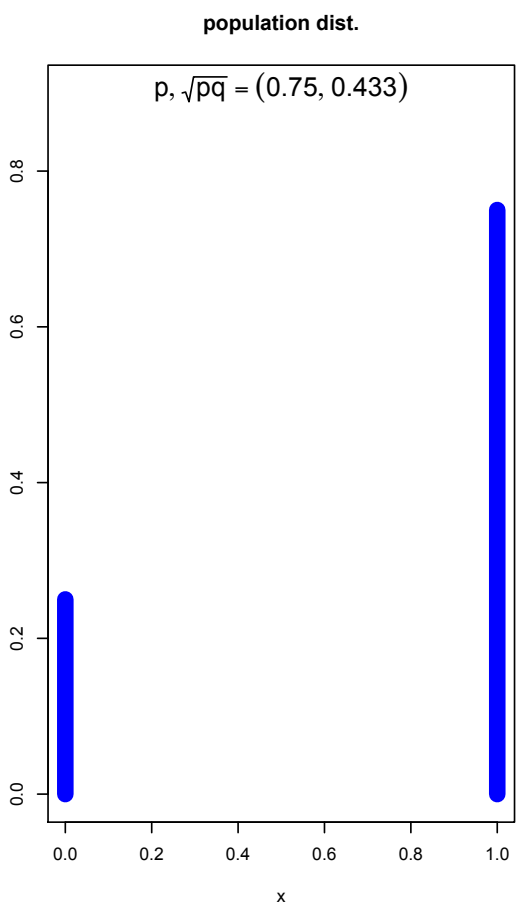
n=50, 母體的p=50%

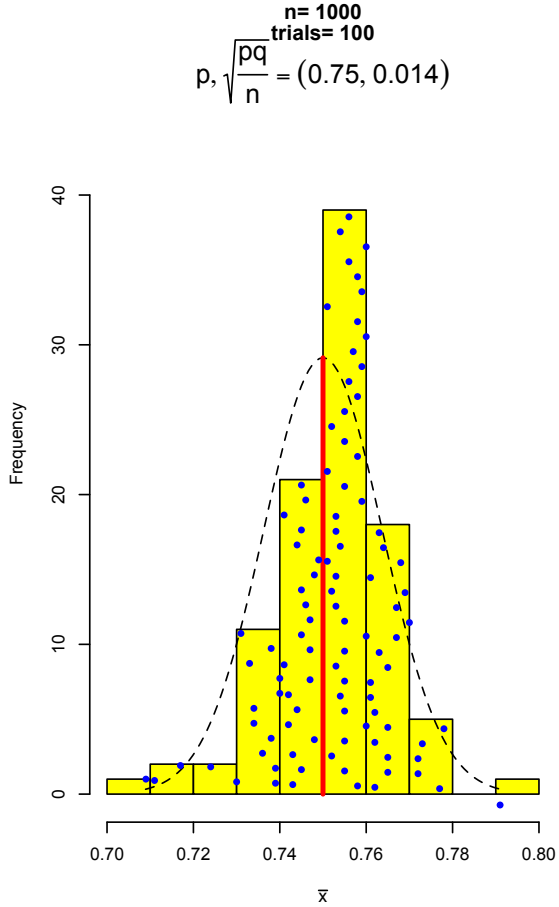

population dist.
p, √pq = (0.5, 0.5)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.2
0.4
0.6
0.8
1.0
x

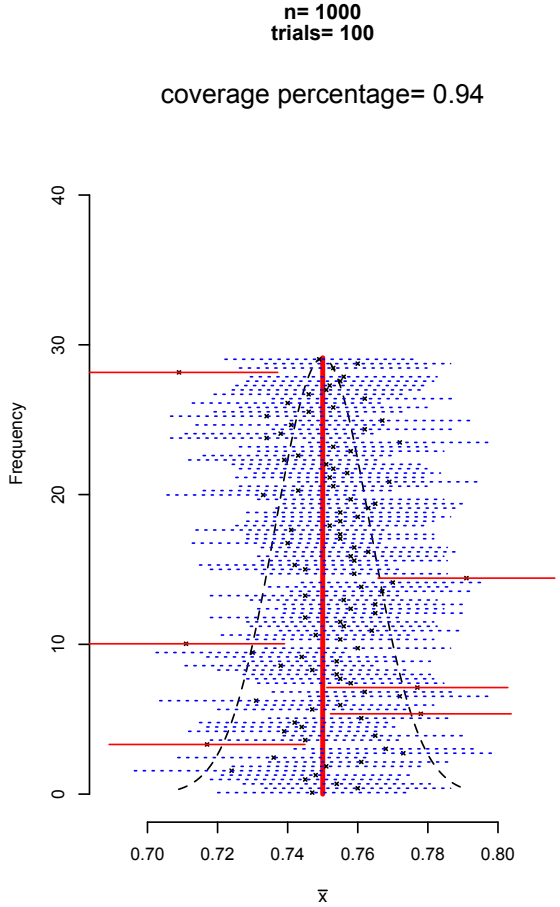

n= 50
trials= 100
p, √(pq/n) = (0.5, 0.071)
Frequency
0
10
20
30
40
0.3
0.4
0.5
0.6
0.7
x̄


n= 50
trials= 100
coverage percentage= 0.94
Frequency
0
10
20
30
40
0.2
0.3
0.4
0.5
0.6
0.7
0.8
x̄

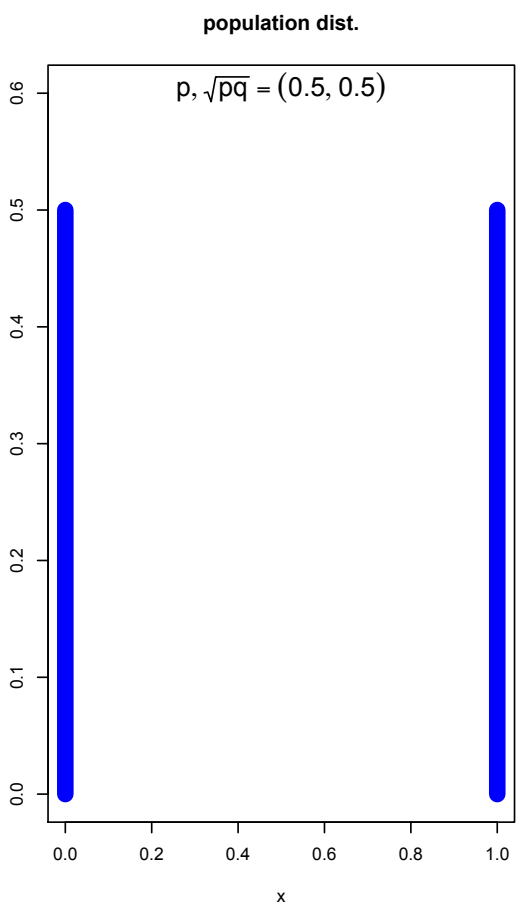
p=75%, n=1000

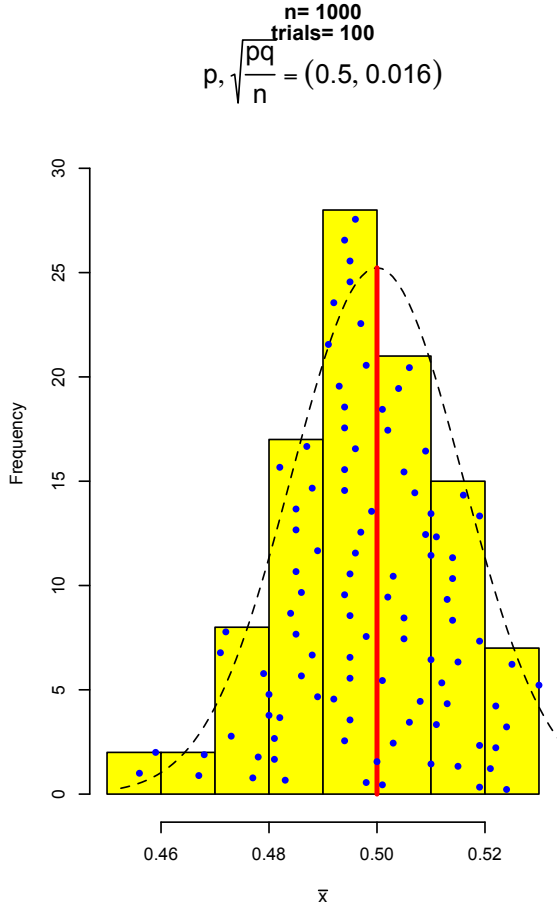

population dist.
p, √pq = (0.75, 0.433)
0.0
0.2
0.4
0.6
0.8
1.0
x

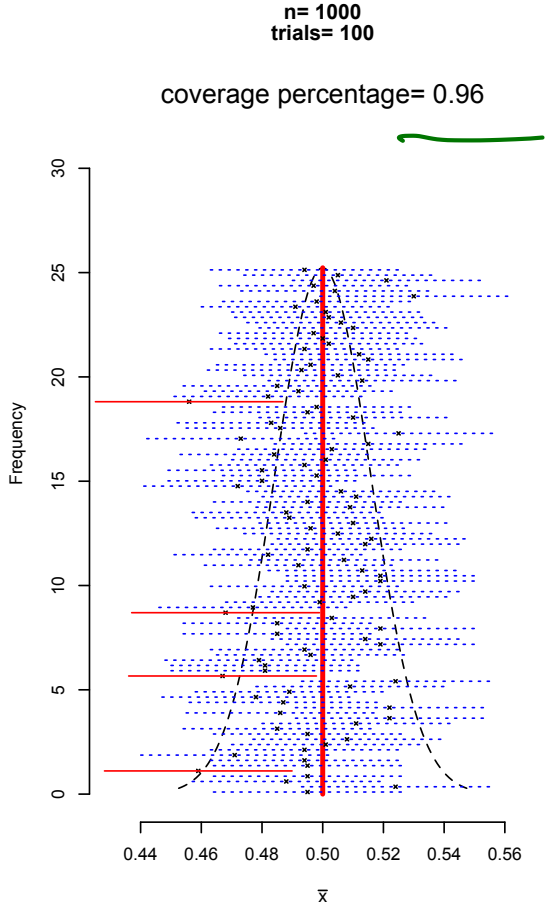

n= 1000
trials= 100
p, √(pq/n) = (0.75, 0.014)
Frequency
0
10
20
30
40
0.70
0.72
0.74
0.76
0.78
0.80
x̄


n= 1000
trials= 100
coverage percentage= 0.94
Frequency
0
10
20
30
40
0.70
0.72
0.74
0.76
0.78
0.80
x̄

p=50%, n=1000


population dist.
p, √pq = (0.5, 0.5)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.2
0.4
0.6
0.8
1.0
x


n= 1000
trials= 100
p, √(pq/n) = (0.5, 0.016)
Frequency
0
5
10
15
20
25
30
0.46
0.48
0.50
0.52
x̄


n= 1000
trials= 100
coverage percentage= 0.96
Frequency
0
5
10
15
20
25
30
0.44
0.46
0.48
0.50
0.52
0.54
0.56
x̄

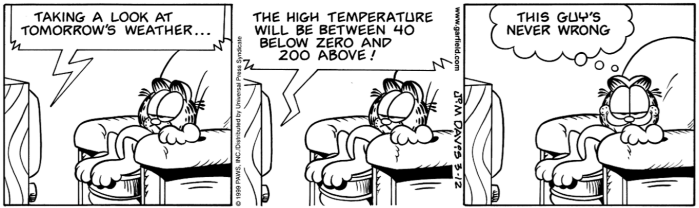
Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

100% CI for p : [0,1]

Can you see any drawbacks to using a wider interval?



If the interval is too wide it may not be informative.

Changing the confidence level

$$point\ estimate \pm z^\star \times SE$$

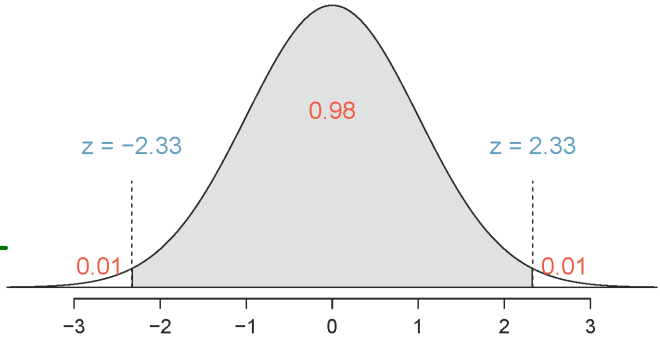
sampling error

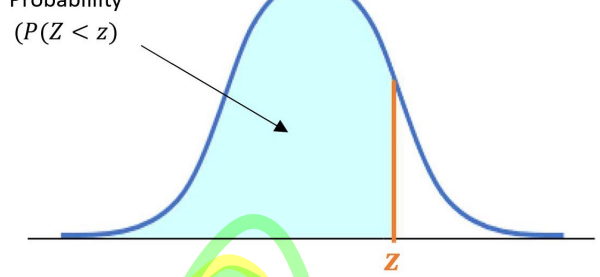
- In a confidence interval, $z^\star \times SE$ is called the margin of error, and for a given sample, the margin of error changes as the confidence level changes.
- In order to change the confidence level we need to adjust $z^\star$ in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, and 99%.
- For a 95% confidence interval, we use $z^\star = 1.96$. Similarly, for confidence levels of 90% and 99%, the corresponding $z^\star$ values are 1.28 and 2.58, respectively.
- However, using the standard normal (z) distribution, it is possible to find the appropriate $z^\star$ for any confidence level.

Which of the below Z scores is the appropriate $z^\star$ when calculating a 98% confidence interval?

(a) $Z = 2.05$
(b) $Z = 1.96$
(c) $Z = 2.33$
(d) $Z = -2.33$
(e) $Z = -1.65$

pnorm
qnorm
dnorm
rnorm





z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5754
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7258	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7518	0.7549
0.7	0.7580	0.7612	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7996	0.8023	0.8051	0.8079	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9485	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9700	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998	0.9998

Interpreting confidence intervals

Confidence intervals are ...

- always about the population
- not probability statements
- only about population parameters, not individual observations

Hypothesis testing framework

Remember when...

Gender discrimination experiment:

		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

$$\hat{p}_{males} = 21/24 \approx 0.88 \text{ and } \hat{p}_{females} = 14/24 \approx 0.58$$

Possible explanations:

- Promotion and gender are independent, no gender discrimination, observed difference in proportions is simply due to chance. $\rightarrow$ null - (nothing is going on)
- Promotion and gender are dependent, there is gender discrimination, observed difference in proportions is not due to chance. $\rightarrow$ alternative - (something is going on)

Recap: hypothesis testing framework

- We start with a null hypothesis (H_0) that represents the status quo.
- We also have an alternative hypothesis (H_A) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or traditional methods based on the central limit theorem (coming up next...).
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

A trial as a hypothesis test

- Hypothesis testing is very much like a court trial.
- H_0: Defendant is innocent
  H_A: Defendant is guilty
- We then present the evidence - collect data.

- Then we judge the evidence - "Could these data plausibly have happened by chance if the null hypothesis were true?"
  - If they were very unlikely to have occurred, then the evidence raises more than a reasonable doubt in our minds about the null hypothesis.
- Ultimately we must make a decision. How unlikely is unlikely?

Image from `http://www.nwherald.com/_internal/cimg!0/oo1il4sf8zzaqbboq25oevvbg99wpot`.

A trial as a hypothesis test (cont.)

- If the evidence is not strong enough to reject the assumption of innocence, the jury returns with a verdict of "not guilty".
  - The jury does not say that the defendant is innocent, just that there is not enough evidence to convict.
  - The defendant may, in fact, be innocent, but the jury has no way of being sure.
- Said statistically, we fail to reject the null hypothesis.
  - We never declare the null hypothesis to be true, because we simply do not know whether it's true or not.
  - Therefore we never "accept the null hypothesis".

- **Burden of Proof:**
  - In trials, the *prosecution* carries the burden of proof.
  - In hypothesis testing, the burden is on the *test statistic*.
- We'll formally introduce the hypothesis testing framework using an example on testing a claim about a proportion.

Testing hypotheses using confidence intervals

Earlier we calculated a 95% confidence interval for the proporton of American Facebook users who think Facebook categorizes their interests accurately as 64% to 67%. Based on this confidence interval, do the data support the hypothesis that majority of American Facebook users think Facebook categorizes their interests accurately.

- The associated hypotheses are:

  H_0: $p = 0.50$: 50% of American Facebook users think Facebook categorizes their interests accurately

  H_A: $p > 0.50$: More than 50% of American Facebook users think Facebook categorizes their interests accurately

- Null value is not included in the interval $\rightarrow$ reject the null hypothesis.
- This is a quick-and-dirty approach for hypothesis testing, but it doesn't tell us the likelihood of certain outcomes under the null hypothesis (p-value).

Decision errors

- Hypothesis tests are not flawless.
- In the court system innocent people are sometimes wrongly convicted and the guilty sometimes walk free.
- Similarly, we can make a wrong decision in statistical hypothesis tests as well.
- The difference is that we have the tools necessary to quantify how often we make errors in statistics.

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	Type 1 Error
	H_A true	Type 2 Error	✓

- A Type 1 Error is rejecting the null hypothesis when H_0 is true.
- A Type 2 Error is failing to reject the null hypothesis when H_A is true.
- We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

P(type I error) ↓ P(type II error) ↓

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0 : \text{ Defendant is innocent}$$
$$H_A : \text{ Defendant is guilty}$$

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

  Type 2 error
- Declaring the defendant guilty when they are actually innocent

  Type 1 error

Which error do you think is the worse error to make?

"better that ten guilty persons escape than that one innocent suffer"
– William Blackstone

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a significance level of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error} \mid H_0 \text{ true}) = \alpha$$

- This is why we prefer small values of α – increasing α increases the Type 1 error rate.

Facebook interest categories

The same survey asked the 850 respondents how comfortable they are with Facebook creating a list of categories for them. 41% of the respondents said they are comfortable. Do these data provide convincing evidence that the proportion of American Facebook users are comfortable with Facebook creating a list of interest categories for them is different than 50%?

$\hat{p} = 0.41$

$H_0 : p = 0.5$

$H_a : p \neq 0.5$

https://www.pewinternet.org/2019/01/16/facebook-algorithms-and-personal-data/

Setting the Hypotheses

- The parameter of interest is the proportion (p) of all American Facebook users who are comfortable with Facebook creating categories of interests for them.
- Initial assumption (Null Hypothesis H_0): 50% of American Facebook users are comfortable with this practice.

$$H_0 : p = 0.50$$

- Alternative Hypothesis (H_A): The actual proportion differs from 50%.

$$H_A : p \neq 0.50$$

$\hat{p} = 0.41$

- Two possible explanations if our sample proportion deviates from 0.50:
  1. The true population proportion significantly differs from 0.50.
  2. The observed difference is due to natural sampling variability, assuming the true population proportion is 0.50.

Test statistic

In order to evaluate if the observed sample proportion is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the test statistic.

H_0 is true, $p = 0.5$

$$\hat{p} \sim N\left(\mu = 0.50, SE = \sqrt{\frac{0.50 \times 0.50}{850}}\right)$$

$$Z = \frac{0.41 - 0.50}{0.0171} = -5.26$$

The sample proportion is 5.26 standard errors away from the hypothesized value. Is this considered unusually low? That is, is the result statistically significant?

Yes, and we can quantify how unusual it is using a p-value.

Facebook Interest Categories - p-value

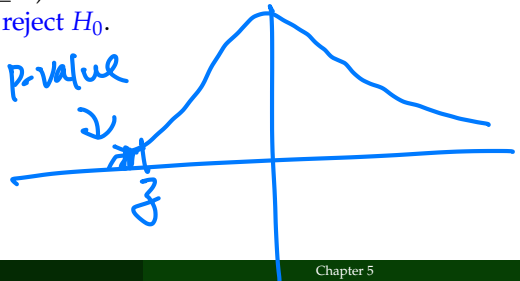
- p-value: The probability of observing data at least as favorable to the alternative hypothesis H_A as our current dataset (a sample proportion less than 0.41), assuming the null hypothesis H_0 (the true population proportion is 0.50) is true.
- Calculation:

$$P(\hat{p} < 0.41 \text{ or } \hat{p} > 0.59 \mid p = 0.50) = P(|Z| > 5.26) < 0.0001$$

This result indicates a very low p-value, suggesting strong evidence against H_0 given our data.

Understanding p-values

- The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, under the assumption that the null hypothesis H_0 is true.
- Low p-value ($< \alpha$, typically 5%): Implies that the observed data is unlikely under H_0, leading us to reject H_0.
- High p-value ($\geq \alpha$): Indicates that the observed data could reasonably occur under H_0, and thus we do not reject H_0.



Facebook interest categories - Making a decision

- p-value < 0.0001
  - If 50% of all American FB users are comfortable with FB creating these interest categories, there is less than a 0.01% chance of observing a random sample of 850 American Facebook users where 41% or fewer or 59% of higher feel comfortable with it.
  - Pretty low probability to think that observed sample proportion, or something more extreme, is likely to happen by chance.
- Since p-value is low (lower than 5%) we reject H_0.
- The data provide convincing evidence that the proportion of American FB users who are comfortable with FB creating a list of interest categories for them is different than 50%.
- The difference between the null value of 0.50 and observed sample proportion of 0.41 is not due to chance or sampling variability.

Choosing a significance level

$\alpha = 0.05$

- The traditional significance level is 0.05, but adjusting it based on context enhances decision-making.
- When the consequences of Type 1 Errors (incorrectly rejecting H_0) are significantly serious, it's prudent to choose a smaller significance level (e.g., 0.01, 0.001), demanding stronger evidence before accepting H_A.
- When Type 2 Errors (failing to reject H_0 when it's false) pose greater risks, a higher level (e.g., 0.10) may be more appropriate to reduce the chance of overlooking a true H_A.

$P(\text{type 1 error}) \leq \alpha$ ($=0.05$)

significance level

One-sided vs. Two-sided Hypothesis Tests and p-values

$H_0: p = 0.5$ $H_a: p < 0.5$ $H_0: p = 0.5$ $H_a: p > 0.5$

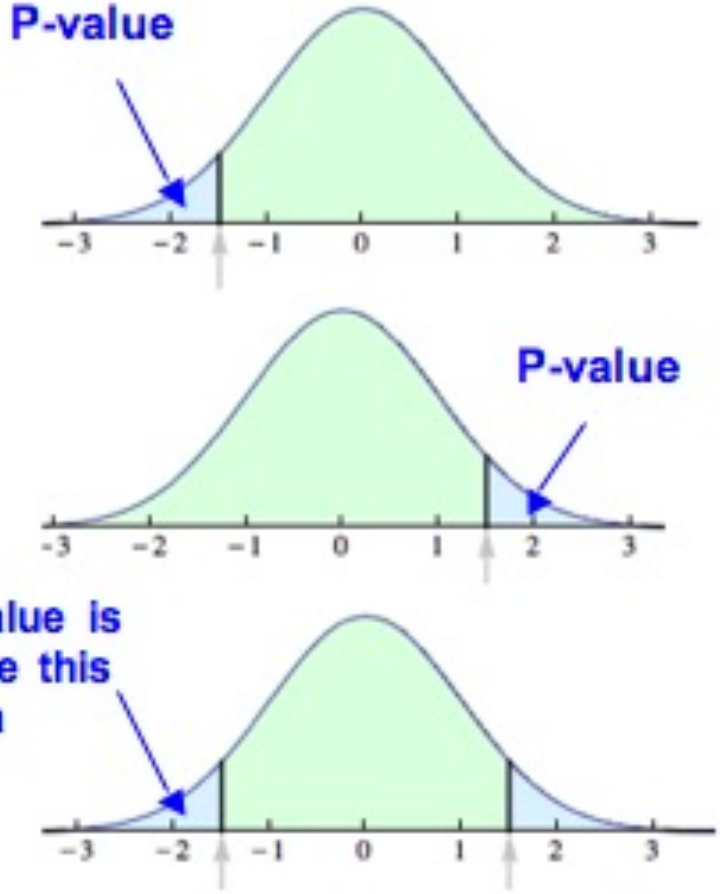
$H_0: p \geq 0.5$ $H_a: p < 0.5$

- In a two-sided test (e.g., whether the proportion of Facebook users comfortable with interest categories is not 50%): $H_A : p \neq 0.50$.
- In a one-sided test, we examine if the deviation is in a specific direction:
  - To check if less than 50% of users are comfortable, use $H_A : p < 0.50$.
- The p-value in a two-sided test is generally twice the p-value of a one-sided test because it accounts for deviations in both directions from the null value.
- Two-sided tests are preferred for a comprehensive analysis, unless prior evidence or theoretical reasons strongly support a one-sided hypothesis.

Two Sided

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

Standard Normal Model



$H_a : p < p_o$

Left-tailed P-value

$H_a : p > p_o$

Right-tailed P-value

$H_a : p \neq p_o$

Two-tailed P-value

①

4/8 : 12:30 – 14:10

4/11 : 13:20 – 14:10

14:20 – 15:10

4/16 Mid-term

13:10 – 14:10

density

dnorm(0) = $\frac{1}{\sqrt{2\pi}}$

$X \sim N(0,1)$ $P(X=0) = 0$

dbinom(1) = $P(X=1)$

prb. fn

pnorm(x) = $P(X \leq x)$