STAT509: Statistical inference for proportion

Peijie Hou

University of South Carolina

October 5, 2014

Overview of Statistical Inference

- From this chapter and on, we will focus on the **statistical inference**.
- Statistical inference deals with making (probabilistic) statements about a population of individuals based on information that is contained in a sample taken from the population.

Terminology: population/sample

- A **population** refers to the entire group of "individuals" (e.g., parts, people, batteries, etc.) about which we would like to make a statement (e.g., defective proportion, median weight, mean lifetime, etc.).
  - Problem: Population can not be measured (generally)
  - Solution: We observe a **sample** of individuals from the population to draw inference
  - We denote a random sample of observations by

$$Y_1, Y_2, \ldots, Y_n$$

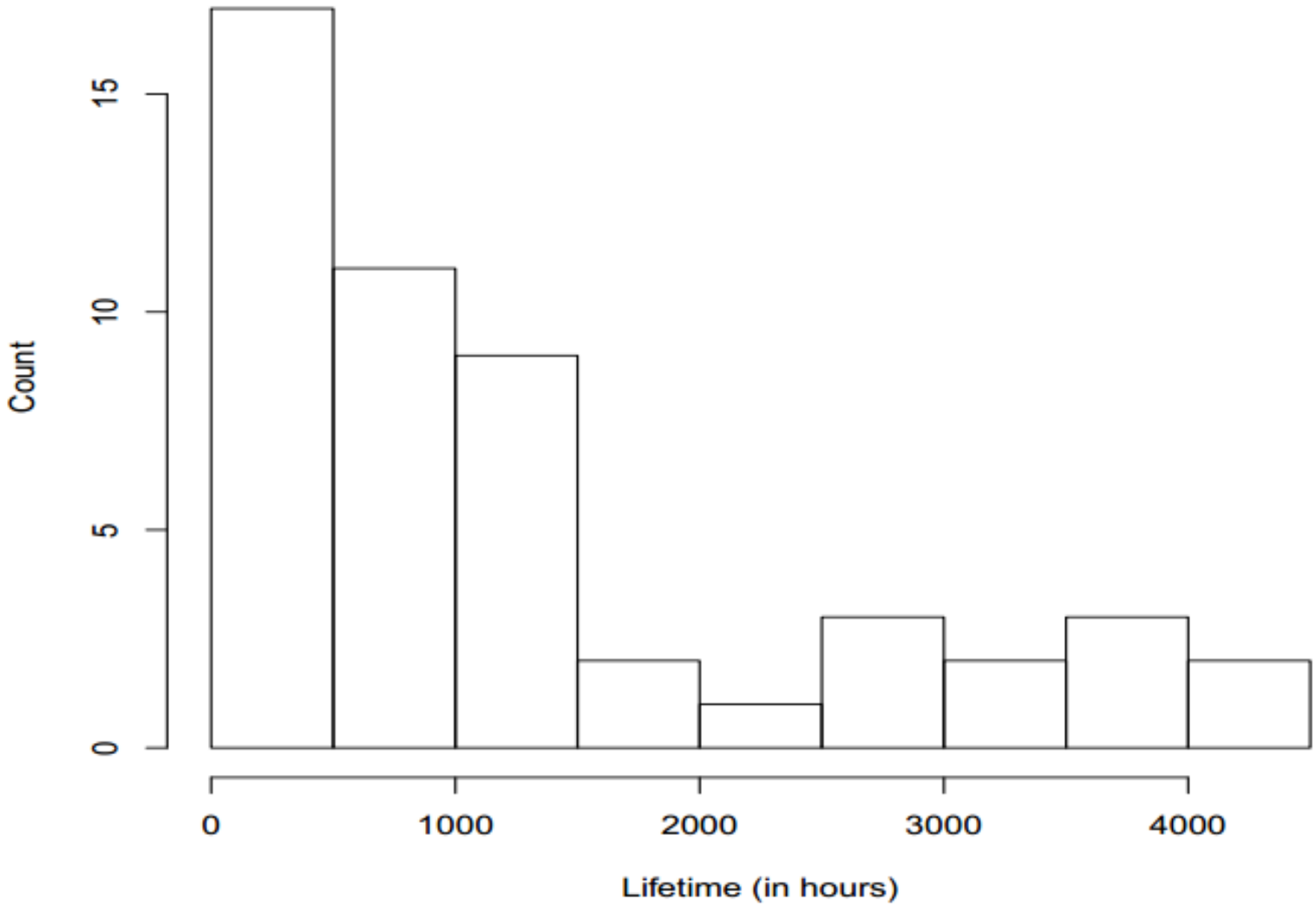
  - n is the **sample size**

Example

BATTERY DATA: Consider the following random sample of $n = 50$ battery lifetimes $y_1, y_2, \ldots, y_{50}$ (measured in hours):

4285	2066	2584	1009	318	1429	981	1402	1137	414
564	604	14	4152	737	852	1560	1786	520	396
1278	209	349	478	3032	1461	701	1406	261	83
205	602	3770	726	3894	2662	497	35	2778	1379
3920	1379	99	510	582	308	3367	99	373	454

A histogram of battery lifetime data



Cont'd on battery lifetime data

The (empirical) distribution of the battery lifetimes is skewed towards the high side

- Which continuous probability distribution seems to display the same type of pattern that we see in histogram?
- An exponential(λ) models seems reasonable here (based in the histogram shape). What is λ?
- In this example, λ is called a (population) **parameter** (generally unknown). It describes the theoretical distribution which is used to model the entire population of battery lifetimes.
- All of the probability distributions that we discussed in previous chapter are meant to describe (model) population behavior.

Terminology: parameter

- A **parameter** is a numerical quantity that describes a *population*. In general, population parameters are unknown.
- Some very common examples are:
  - μ = population mean
  - σ^2 = population variance
  - σ = population standard deviation
  - p = population proportion
- Connection: all of the probability distributions that we talked about in previous chapter are indexed by population (model) parameters.

Terminology: statistics

- A **statistic** is a numerical quantity that can be calculated from a sample of data.
- Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from a population, some very common examples are:
  - **sample mean**:
$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$
  - **sample variance**:
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$
  - **sample standard deviation**: $S = \sqrt{S^2}$
  - **sample proportion**: $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ if $Y_i's$ are binary.

Back to battery lifetime data

With the battery lifetime data (a random sample of $n = 50$ lifetimes),

$$\bar{y} = 1274.14 \text{ hours}$$
$$s^2 = 1505156 \text{ (hours)}^2$$
$$s \approx 1226.85 \text{ hours}$$

R code:

```
> mean(battery) ## sample mean
 [1] 1274.14
> var(battery) ## sample variance
 [1] 1505156
> sd(battery) ## sample standard deviation
 [1] 1226.848
```

Parameters and Statistics Cont'd

SUMMARY: The table below succinctly summarizes the salient differences between a population and a sample (a parameter and a statistic):

Comparison between parameters and statistics	
Statistics	*Parameters*
• Describes a sample • Always known • Changes upon repeated sampling • Ex: $\bar{X}, S^2, S$	• Describes a population • Usually unknown • Is fixed but unknown • Ex: μ, σ^2, σ

Point estimators and sampling distributions

- Let θ denote a population parameter.
- A **point estimator** $\hat{\theta}$ is a statistic that is used to estimate a population parameter θ.
- Common examples of point estimators are:
  - $\widehat{\theta} = \overline{Y} \longrightarrow$ a point estimator for $\theta = \mu$
  - $\widehat{\theta} = S^2 \longrightarrow$ a point estimator for $\theta = \sigma^2$
  - $\widehat{\theta} = S \longrightarrow$ a point estimator for $\theta = \sigma$
- Remark: In general, $\widehat{\theta}$ is a statistic, the value of $\widehat{\theta}$ will vary from sample to sample.

Terminology: sampling distribution

- The distribution of an estimator $\widehat{\theta}$ is called its **sampling distribution**.
- A sampling distribution describes mathematically how $\widehat{\theta}$ would vary in repeated sampling.
- What is a good estimator? And good in what sense?

Evaluate an estimator

- **Accuracy:** We say that $\widehat{\theta}$ is an **unbiased estimator** of θ if and only if
$$E(\widehat{\theta}) = \theta$$
- *RESULT:* When $Y_1, \ldots, Y_n$ is a random sample,
$$E(\overline{Y}) = \mu$$
$$E(S^2) = \sigma^2$$
- **Precision:** Suppose that $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are unbiased estimators of θ. We would like to pick the estimator with smaller variance, since it is more likely to produce an estimate close to the true value θ.

Evaluate an estimator: cont'd

- *SUMMARY:* We desire point estimators $\widehat{\theta}$ which are **unbiased** (perfectly accurate) and have **small variance** (highly precise).
- *TERMINOLOGY:* The **standard error** of a point estimator $\widehat{\theta}$ is equal to

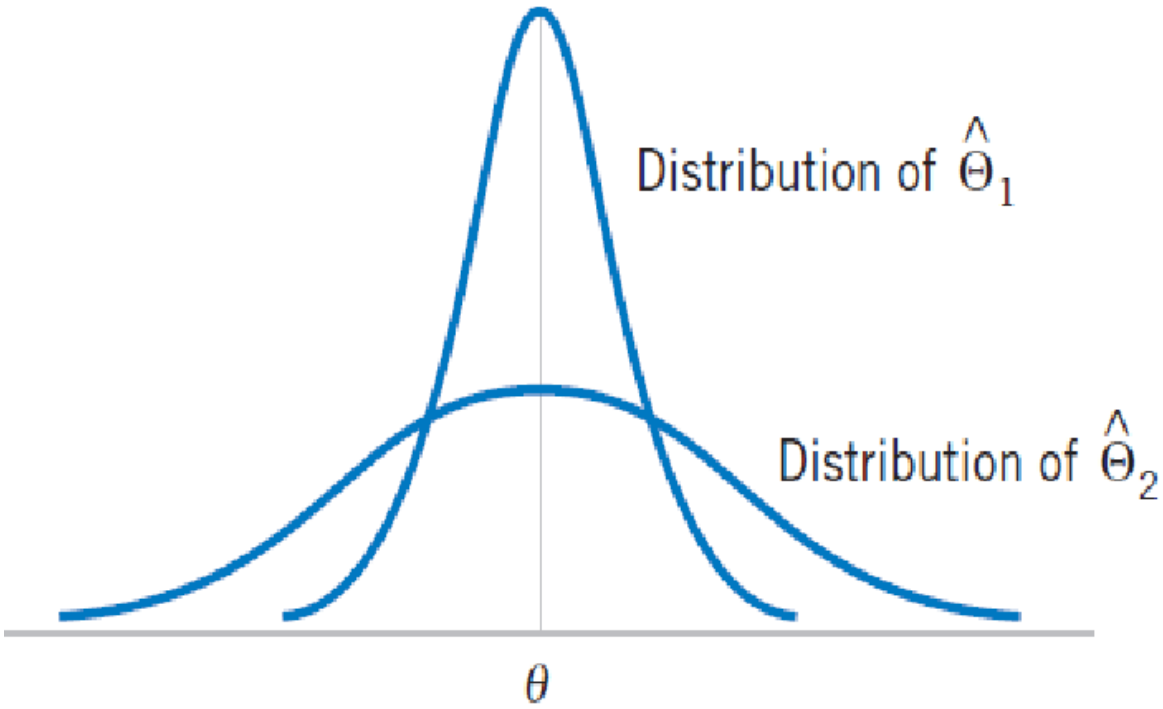
$$se(\widehat{\theta}) = \sqrt{\text{var}(\widehat{\theta})}.$$

- Note:

$$\text{smaller } se(\widehat{\theta}) \Longleftrightarrow \widehat{\theta} \text{ more precise.}$$

Evaluate an estimator: cont'd

Which estimator is better? Why?



Inference on Population Proportion

The population proportion p emerges when the characteristic we measure on each individual is binary (i.e., only 2 outcomes possible). Here are some examples:

p = proportion of airline has experienced exceedence

p = proportion of defective water filters in a factory

p = proportion of HIV positive in SC

We can connect these binary outcomes to the *Bernoulli trails* assumptions for each individual in the sample:

1. Each trial results in only two possible outcomes, labeled as "success" and "failure."
2. The trials are independent.
3. The probability of a success in each trial, denoted as p, remains constant. It follows that the probability of a failure in each trial is $1 - p$.

Point Estimator of Proportion p

- Suppose we define $Y =$ the number of successes out of n sampled individuals so $Y \sim b(n, p)$. A natural point estimator for p, *the population proportion*, is

$$\widehat{p} = \frac{Y}{n},$$

the **sample proportion**. $\widehat{p}$ is read as p hat.

Property of $\widehat{p}$

- $\widehat{p}$ is a unbiased estimator of p. That is,

$$E(\widehat{p}) = p.$$

- To quantify the precision of $\widehat{p}$,

$$\mathrm{var}(\widehat{p}) = \frac{p(1-p)}{n}$$

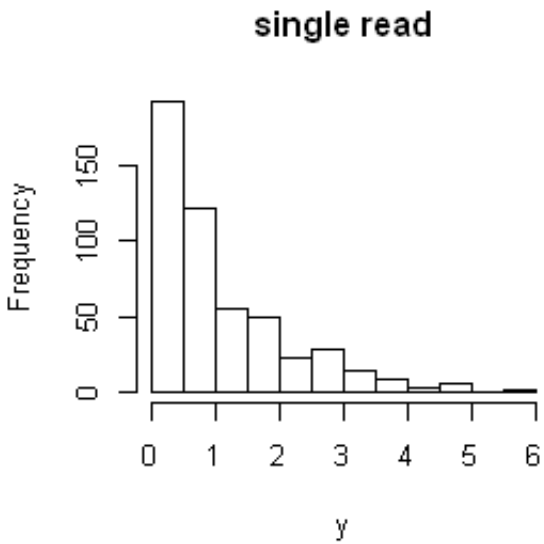
- Question: What is the (asymptotic) distribution of $\widehat{p}$?

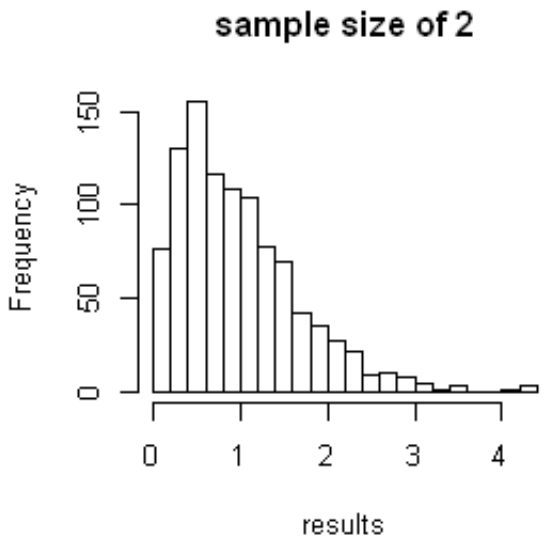
Sampling Distribution of $\widehat{p}$ and CLT

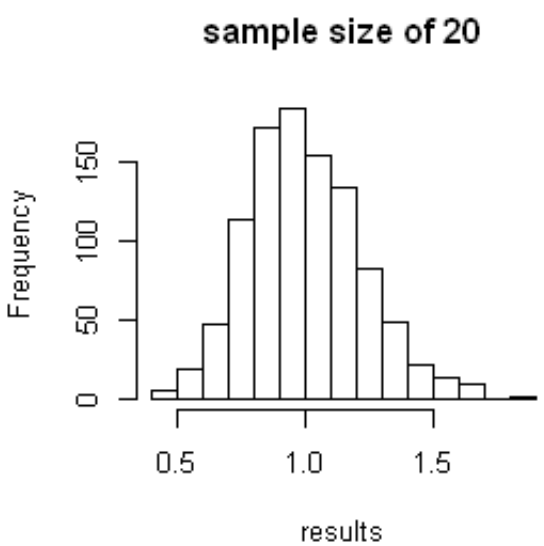
- To derive the sampling distribution of $\widehat{p}$, we need first introduce the central limit theorem.
- Central Limit Theorem: Suppose that $Y_1, Y_2, \ldots, Y_n$ is a random sample from a population distribution with mean μ and variance σ^2. When the sample size n is large, we have

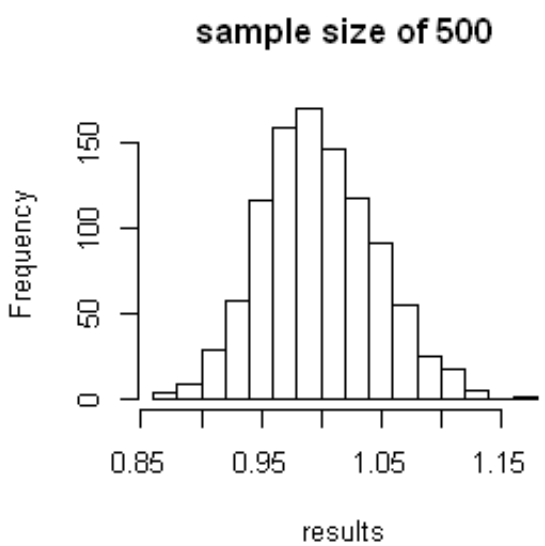
$$\overline{Y} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right)$$

Simulation Study of CLT Cont'd









Sampling distribution of $\hat{p}$

With the help of central limit theorem, we can derive an asymptotic distribution of $\hat{p}$. Recall that $Y =$ the number of successes out of n sampled individuals and $Y \sim b(n, p)$. We can express Y as the sum of n independent Bernoulli trials with success probability p. That is

$$Y = \sum_{i=1}^{n} X_i,$$

where $X \overset{i.i.d.}{\sim}$ Bernoulli(p). $E(X_i) = p$, and $\text{Var}(X_i) = p(1-p)$. It follows that $\widehat{p} = Y/n = \sum_{i=1}^{n} X_i/n = \bar{X}$. By CLT, we have

$$\widehat{p} = \frac{Y}{n} = \overline{X} \sim \mathcal{AN}\left(p, \frac{p(1-p)}{n}\right).$$

Confidence Interval

- Using a point estimator only **ignores important information**; namely, how variable the estimator is.
- To avoid this problem (i.e., to account for the uncertainty in the sampling procedure), we therefore pursue the topic of interval estimation (also known as confidence intervals).
- The main difference between a point estimate and an interval estimate is that
  - a **point estimate** is a one-shot guess at the value of the parameter; this ignores the variability in the estimate.
  - an **interval estimate** (i.e., **confidence interval**) is an interval of values. It is formed by taking the point estimate and then adjusting it downwards and upwards to account for the point estimate's variability.

Confidence Interval for p (p.270 in textbook)

- Recall that $\widehat{p} \sim \mathcal{AN}\left(p, \frac{p(1-p)}{n}\right)$.
- Let us define z_α be the upper α percentage point of the standard normal distribution, i.e., $P(Z > z_\alpha) = \alpha$.
- We will derive a $100(1-\alpha)\%$ CI for p on blackboard:

Confidence Interval for p cont'd

- An approximate $100(1-\alpha)\%$ confidence interval for p is

$$\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \quad \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right).$$

- The quantity $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is called the **margin of error** .
- Rule of thumb: To use normal approximation, we need $np \geq 15$ and $n(1-p) \geq 15$.
- Note of the form of the interval:

$$\text{point estimate} \ \pm \ \text{quantile} \times \text{standard error}$$

Interpretation of Confidence Interval

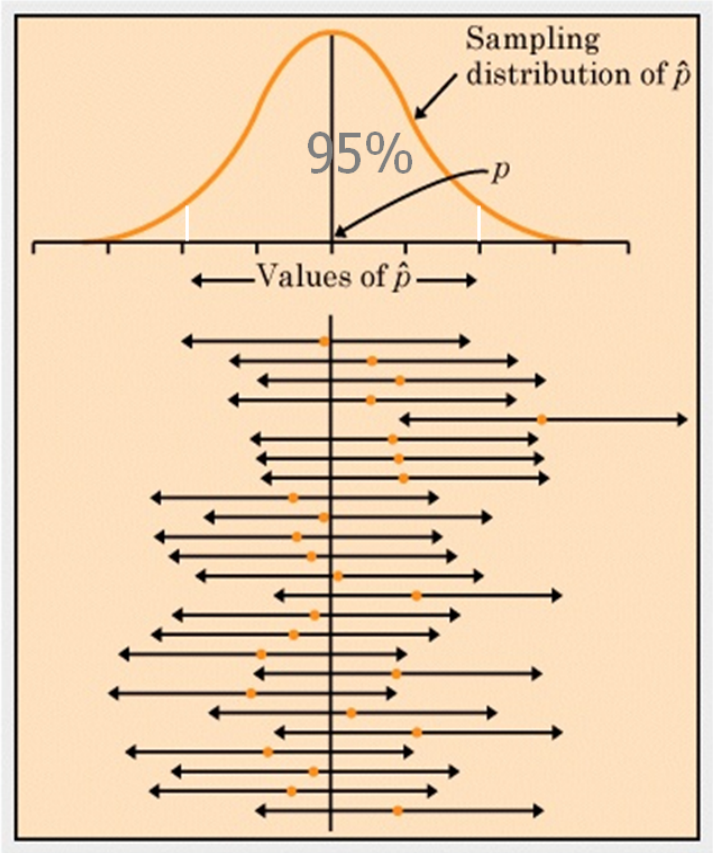
- Suppose that we are interested in parameter p for certain population. We take a sample of size n and calculate the sample proportion $\hat{p} = Y/n$. A 95% confidence interval is given by

$$\text{Point. Est} \pm 1.96 \text{ Standard Error.}$$

- The 95% confidence comes from the fact that if we repeated this experiment over and over again, apprxomiately, 95% of all samples would produce a confidence interval that contains the true proportion, and only 5% of the time would the interval be in error.
- We call $100(1-\alpha)\%$ the **confidence level**.

Interpretation of Confidence Interval Cont'd

Here is a pictorial illustration of the confidence interval:



Statistical Hypothesis

- Definition: a *statistical hypothesis* is an assertion or conjecture concerning one or more population parameters.
- Example:
  1. The proportion of underweight milk is more then 3% in a local farm.
  2. More than 7% of the landings for a certain airline exceed the runway.
  3. The defective rate of the water filter is less than 5%.

4 Steps to a Hypothesis Test

1. State the **null** and **alternative** hypotheses.
2. Collect the data and calculate **test statistic** assuming H_0 is true.
3. Assuming the null hypothesis is true, calculate the p-**value**.
4. Draw conclusion based on the p-value. We either **reject** H_0 or **fail to reject** H_0.

Let us look at an example to illustrate these steps...

Example: Defective Water Filters

- Historically, the defective rate of water filters is 7% in a certain factory. A new quality control system is introduced to reduce the defective rate. Suppose that we randomly choose 300 water filters, and calculate $\hat{p} = 0.041$. We want to test whether the new system reduce the defective rate or not.
- Let p=proportion of defective water filters in the factory after introducing the new system.

Step 1: The Null and Alternative Hypothesis

- *Null hypothesis* is denoted by H_0, which represents what we assume to be true. Under null hypothesis, the exact value of the parameter is specified.
- *Alternative hypothesis* is denoted by H_a, which represents the researcher's interest.
- In most situation, researchers want to reject null hypothesis in favor of the alternative hypothesis by performing some experiment.
- In defective water filters example,

$$H_0 : p = 0.07$$
$$H_a : p < 0.07 \quad \text{(the new system reduces the defective rate)}$$

Step 2: Calculate test statistic

- How should we make our decision based on the sample?
- We reject H_0, if $\widehat{p}$ is far less than 0.07, which is not likely to happen if assuming $p = 0.07$.
- We need the sampling distribution to quantify how far is far.

Test Statistic for Proportion

- Recall that if $H_0 : p = p_0$ is true, then

$$\widehat{p} \sim \mathcal{AN}\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right).$$

- Therefore, the test statistic is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim \mathcal{AN}(0,1)$$

- In defective water filter example, assuming H_0 is true, the test statistic is calculated as

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.041 - 0.07}{\sqrt{\frac{0.07(1-0.07)}{300}}} = -1.97.$$

- Question: is this number likely to appear if assume $p_0 = 0.07$ is true?

Step 3: Calculate p-value

- The *p*-**value** is the probability of getting the sample results you got or something more extreme assuming that the null hypothesis is true.
- If the *p*-value is small, we doubt the null hypothesis since it is not likely to observe such a "extreme" test statistic under H_0. There is evidence to against null hypothesis.
- On the other hand, if the *p*-value is large, we have a pretty good chance to observe the computed test statistic under H_0 in a single experiment, there is no reason to question the H_0.
- In other words, the p-value for a hypothesis test measures how much evidence we have against H_0, that is,

$$\text{the smaller the p-value} \Longrightarrow \text{the more evidence against } H_0$$

Step 3: Calculate p-value cont'd

- In defective water filter example, the p-value of the test is:

$$P(Z < -1.97) = 0.024 \qquad \text{(found in table)}$$

- Why " $<$ "?

Alternative hypothesis	Hypothesis type	p-value formula
$H_a : p < p_0$	Left-tail hypothesis	$\mathrm{P}(Z < z_0)$
$H_a : p > p_0$	Right-tail hypothesis	$\mathrm{P}(Z > z_0)$
$H_a : p \neq p_0$	Two-tail hypothesis	$2\mathrm{P}(Z < -\lvert z_0 \rvert)$

- In our example, the p-value is 0.024, do you think it is large?

Step 4: Conclusion

- If the p-value is **small**, we **reject the null hypothesis** and **conclude the alternative hypothesis**.
- If the p-value is **not small**, we **do not reject the null hypothesis** and **do not conclude the alternative hypothesis.**
- There is one remaining question, how small should p-value be to be considered as "small"? We need level of significance to answer it.

Step 4: Conclusion cont'd

- We use α to denote the level of significance.
- Level of significance is determined before you see the data.
- In practice, we usually set $\alpha = 0.05$. Other common choices are $\alpha = 0.01$, or $\alpha = 0.1$.
- We simply compare the p-value with the α level, we reject H_0 if the p-value is less than α; and do not reject H_0 is the p-value is greater than or equal to α
- In defective water filter example, p-value$= 0.024 < 0.05$ (pre-defined), therefore, we reject H_0, and conclude that we have sufficient evidence to conclude the new system reduces the defective rate (note: we conclude H_a in the context of the question.)

Example: Exceedance of the Localizer

A certain type of flu outbreaks in northern part of the USA. The historical records shows that there are 7% of the residences in Columbia carrying flu under usual condition. Researchers want to see whether there is an outbreak in Columbia, there are 30 out of 250 randomly chosen people in the sample carrying flu. Can we conclude that there is an outbreak in Columbia? Answer the following question using confidence interval approach (95%) and hypothesis testing approach (assuming level of significance is 0.05).

Inference for p: Confidence Interval Approach

Recall that a $100(1-\alpha)\%$ C.I. is

$$\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \quad \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right).$$

- The point estimate $\widehat{p} = 30/250 = 0.12$.
- $z_{\alpha/2} = z_{0.025} = 1.96$
- Standard error: $\sqrt{0.12(1-0.12)/250} = 0.021$
- 95% CI is: (0.079,0.161)
- Conclusion?

Inference for p: Hypothesis Test

- *Step 1: State H_0 and H_1*
  $H_0 : p = 0.07$
  $H_a : p > 0.07$
- *Step 2: Calculate test statistic assuming H_0 is true*

$$Z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.12 - 0.07}{\sqrt{\frac{0.07(1-0.07)}{250}}} = 3.10.$$

- *Step 3: Calculate p-value*
  p-value$=P(Z > 3.10) \approx 0.001$.
- *Step 4: Dawn the conclusion*
  $\alpha = 0.05$, p-value is smaller than 0.05. We reject H_0, and conclude that there is an outbreak in Columbia.

Method of Evaluating a Test: Type I and Type II Errors

There are two mistakes we can make in a hypothesis test.

- Type I error: H_0 is rejected but in reality H_0 is true
- Type II error: H_0 is not rejected but in reality H_0 is false

	Do not reject H_0	Reject H_0
H_0 is true	No Error	Type I error
H_0 is false	Type II error	No Error

Controlling Risk

- The probability of type I error is denoted by α (same as the level of significance), i.e.,

$$\alpha = \text{Prob}\,(\text{reject } H_0 | H_0 \text{ is true}).$$

- The type II error is denoted by β, i.e.,

$$\beta = \text{Prob}\,(\text{fail to reject } H_0 | H_a \text{ is true at some value}).$$

- The idea situation is both type I and type II error is 0, which means we can always make correct decision. However, only oracle knows the true. For us, every decision we make will have associated error probability.

Controlling Risk Cont'd

- In practice, if we try to decrease the type I error, the type II error will increase, and vice versa. Remember, there is no free lunch!
- Researchers should consider the consequences of type I error and type II errors to help determine significance level.
- Example: An environmentalist takes samples at a nearby river to study the average concentration level of a contaminant. He wants to find out, using a 0.1 level of significance, if the average concentration level exceeds the acceptable level for safely consuming fish from the river.

Controlling Risk Cont'd

- Describe a Type I error for this problem and the potential consequence.

- Describe a Type II error for this problem and the potential consequence.