Statistics for Business and Machine learning
Part 2: Inferential Statistics

Tien-Nam Le

tien-nam.le@ens-lyon.fr

Sciences U University
18 Octobre, 2018

Outline

Part II. Inferential Statistics

Sampling

1. Central limit theorem
2. Sampling proportion

Point estimators

1. Unbiased estimators
2. Estimating mean and variance

Confidence intervals

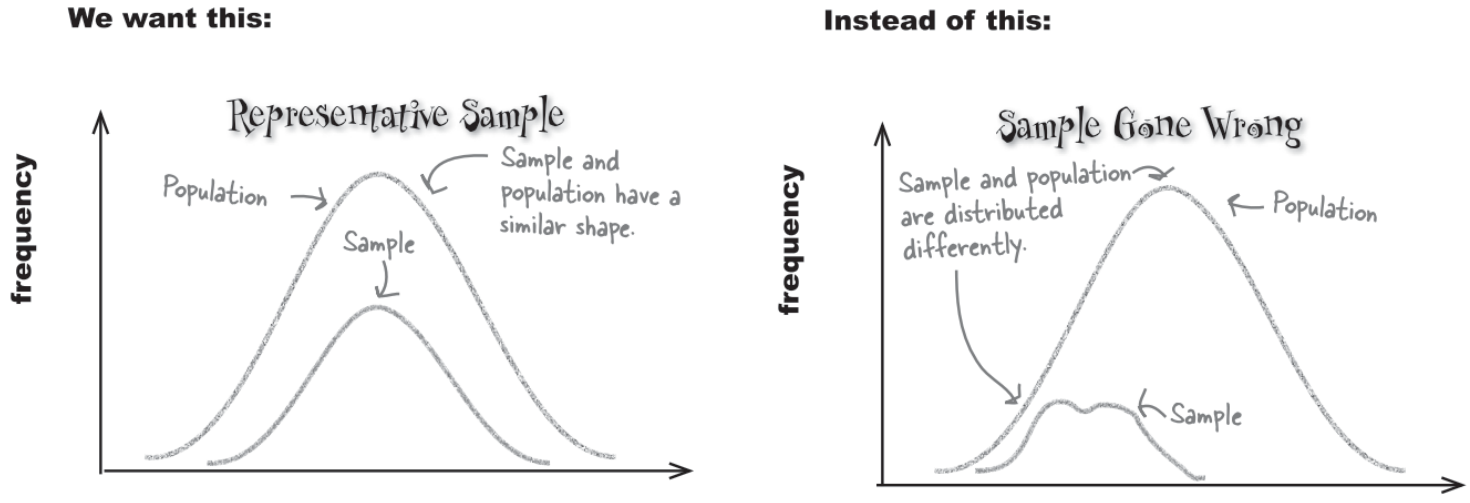
1. Estimating mean when variance is known
2. Estimating mean when variance is unknown

Hypothesis testing

1. significance level, p-value
2. Population proportion

Sampling

Sampling



From now on, a sample $X = \{X_1, ..., X_n\}$ always means:

- Each X_i is independent from other
- Each X_i is random following the distribution of the population

This method of sampling is proved to be good.

Summary

Recall the key parameters:

- Population
  - mean: μ
  - variance: σ^2
  - standard deviation: σ
- Sample
  - mean: $\overline{X} = \dfrac{\sum_i X_i}{n}$
  - variance: $s^2 = \dfrac{\sum (X_i - \overline{X})^2}{n-1}$
  - standard deviation: $s = \sqrt{\dfrac{\sum (X_i - \overline{X})^2}{n-1}}$

The objective is always: knowing some parameters how to find or approximate some other parameters.

Sample mean

The sample mean $\overline{X}$ is a random variable with

- Expected value:

$$\mathbb{E}[\overline{X}] = \mathbb{E}\left[\frac{\sum_i X_i}{n}\right] = \frac{1}{n}\sum_i \mathbb{E}[X_i] = \frac{1}{n} n\mu = \mu$$

- Variance: $Var(\overline{X}) = \dfrac{\sigma^2}{n} \to 0$ when $n \to \infty$.
- Standard deviation: $\sigma(\overline{X}) = \dfrac{\sigma}{\sqrt{n}}$.

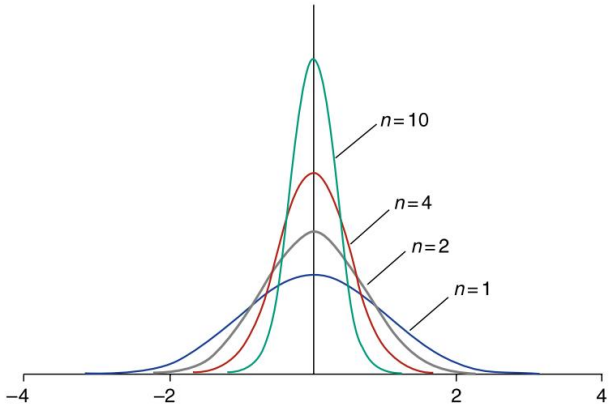


FIGURE 7.1
Densities of sample means from a standard normal population.

Note: The variance above
is variance of sample mean ($Var(\overline{X})$, a number),
not sample variance (s^2, a random variable).

Exercise: The amount of money withdrawn in each transaction at an automatic teller of a branch of the Bank of America has mean $80 and standard deviation $40. What are the mean and standard deviation of the average amount withdrawn in the next 20 transactions?

Central limit theorem

- **Central Limit Theorem:** when n is large ($n \geq 30$), $\overline{X}$ follows normal law with mean μ and standard deviation $\sigma/\sqrt{n}$.
- This implies

$$\mathbb{P}(\overline{X} \leq a) = \mathbb{P}\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{a - \mu}{\sigma/\sqrt{n}}\right) \approx \mathbb{P}\left(Z \leq \frac{a - \mu}{\sigma/\sqrt{n}}\right),$$

  where $Z \sim \mathcal{N}(0, 1)$.
- Similar for $\mathbb{P}(\overline{X} \geq a)$ and $\mathbb{P}(a \leq \overline{X} \geq b)$.
- Calculate probability of Z at https://www.mathsisfun.com/data/standard-normal-distribution-table.html

Exercise: An insurance company has 10,000 automobile policyholders. If the expected yearly claim per policyholder is \$260 with a standard deviation of \$800, approximate the probability that the total yearly claim exceeds \$2.8 million.

Exercise: The blood cholesterol levels of a population of workers have mean 202 and standard deviation 14. If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.

Special case: sampling proportion

This is a very common special case: when the population choose either A or B.

- the population proportion (percentage) choosing A is p and B is $1 - p$ (p may be unknown).
- Sampling proportion $\overline{X}$: proportion of a sample choosing A.
- $\overline{X}$ follows binomial law with expected value p and variance $\frac{p(1-p)}{n}$
- And $\overline{X}$ can be approximated by normal law as in previous slide.

Exercise: Suppose that 46 percent of the population is planning on voting for candidate A in an upcoming election. If a random sample of size 200 is chosen, then

(a) *what is the expected value and standard deviation of the proportion of those in the sample who favor candidate A ?*

(b) *what is the probability that at least 100 favor candidate A?*

More exercises

Exercise: An advertising agency ran a campaign to introduce a product. At the end of its campaign, it claimed that at least 25 percent of all consumers were now familiar with the product. To verify this claim, the producer randomly sampled 1000 consumers and found that 232 knew of the product. If 25 percent of all consumers actually knew of the product, what is the probability that as few as 232 (that is, 232 or less) in a random sample of 1000 consumers were familiar?

Exercise: Suppose that 12 percent of the members of a population are left-handed. In a random sample of 100 individuals from this population,

- *Find the mean and standard deviation of the number of left- handed people.*
- *Find the probability that this number is between 10 and 14 inclusive.*

Exercise: If X is binomial with parameters $n = 80$ and $p = 0.4$, approximate the following probabilities.

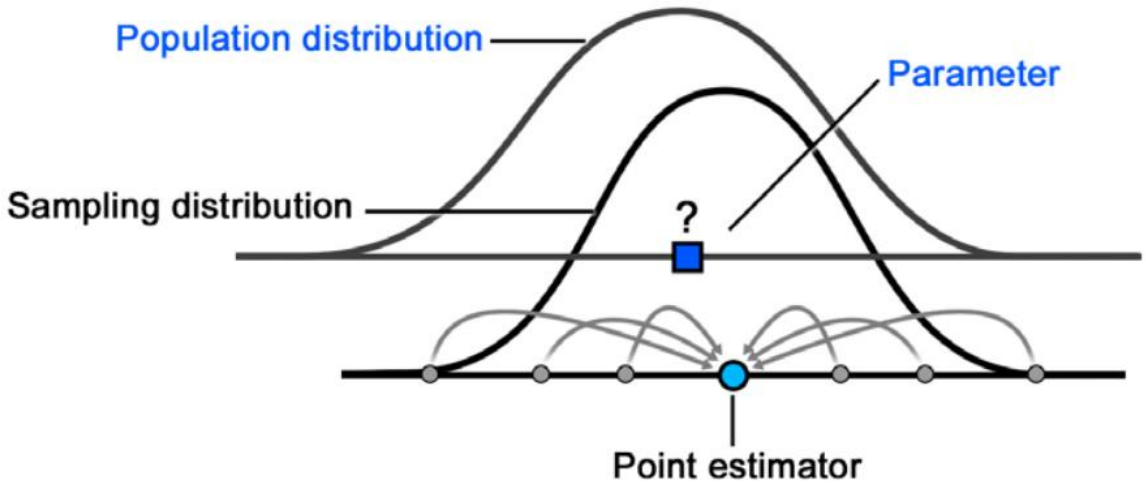
- $\mathbb{P}(X > 34)$
- $\mathbb{P}(X \leq 42)$
- $\mathbb{P}(25 \leq X \leq 39)$

Point Estimator

Point Estimator

Objective: approximate some population parameter such as population mean μ, population variance σ^2.

- *Estimateur* (*point estimator*): approximate a population parameter θ by some parameter obtained from sample.
- Example: To estimate moyenne μ de la population: We have sample $X = \{X_1, ..., X_n\}$, we can choose sample mean $\overline{X}$ as a point estimator.



Point Estimator

Note: There are many point estimators for a population parameter.

Example: To estimate moyenne μ de la population, we can choose:

- $\hat{\mu} = \overline{X}$, moyenne de la donnée
- $\hat{\mu} = \mathit{median}(X)$, médiane de la donnée
- $\hat{\mu} = \dfrac{\max(X) + \min(X)}{2}$

Note: Some estimators are good, some are bad.

Question: How to choose a good parameter?

Biais

- Un estimateur $\hat{\theta}$ is a random variable. Its value varies from sample to sample.
- Le *biais* (*bias*) de l'estimateur $\hat{\theta}$ est

$$\textit{biais}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

- Un estimateur $\hat{\theta}$ est *sans biais* (*unbiased*) if $\mathbb{E}[\hat{\theta}] = \theta$, i.e. $\textit{biais}(\hat{\theta}) = 0$.
- The standard deviation $\sigma(\hat{\theta})$ is usually called the *Erreur type* (*standard error*) of $\hat{\theta}$.
- Un estimateur sans biais $\hat{\theta}$ est *consistent* (*consistent*) si $\sigma(\hat{\theta}) \to 0$ (également $\textit{Var}(\hat{\theta}) \to 0$) quand la taille de l'échantillon $\to \infty$.

Exercises on estimating population mean μ

Exercise: Given $X = \{X_1, ..., X_n\}$ where X_i are randomly chosen from population. Is $\overline{X}$ an unbiased estimator of μ ? What is the standard error of $\overline{X}$? Is $\overline{X}$ a consistent estimator?

Exercise: A proposed study for estimating the average cholesterol level of working adults calls for a sample size of 1000. If we want to reduce the resulting standard error by a factor of 4, what sample size is necessary?

Exercise: A survey is being planned to discover the proportion of the population that is in favor of a new school bond. How large a sample is needed in order to be certain that the standard error of the resulting estimator is less than or equal to 0.1 ?

Estimating population variance σ^2

Objective: Find an estimator for population variance σ^2. Depends on whether you know population mean μ or not.

- If the population mean μ is *known*, then use $\dfrac{\sum(X_i - \mu)^2}{n}$, which is an unbiased estimator.
- If the population mean μ is *unknown*, then use sample variance $s^2 = \dfrac{\sum(X_i - \overline{X})^2}{n-1}$, which is an unbiased estimator.
- Never use $\dfrac{\sum(X_i - \overline{X})^2}{n}$, because it is a biased estimator.

Exercises

Exercise: The following data refer to the amounts (in tons) of chemicals produced daily at a chemical plant.

$$776, 810, 790, 788, 822, 806, 795, 807, 812, 791$$

Use them to estimate the mean and the variance of the daily production.

Exercise: Consistency is of great importance in manufacturing baseballs, for one does not want the balls to be either too lively or too dead. The balls are tested by dropping them from a standard height and then measuring how high they bounce. A sample of 30 balls resulted in the following summary statistics:

$$\sum_{i}^{30} X_i = 52.1 \quad \textit{and} \quad \sum_{i}^{30} X_i^2 = 136.2.$$

Estimate the standard deviation of the size of the bounce.

Intervalle de Confiance

Intervalle de confiance

- Instead of estimating a parameter by a point, we use an interval instead.
- The typical interval estimator of θ is $(\hat{\theta} - c, \hat{\theta} + c)$ for some c.
- The interval has *confiance* (*confidence*) 95% if the interval $(\hat{\theta} - c, \hat{\theta} + c)$ contains θ with probability 95%.
  Note: It doesn't mean $(\theta - c, \theta + c)$ contains $\hat{\theta}$ with probability 95%.
- The length of the interval in this case is $2c$
- c is chosen by the confidence level (95% is default)
- If we want to increase 95% (to say 99%) then c must be increased, i.e. the interval is longer

Case 1. Estimate mean μ when variance σ^2 is known

Theorem: Suppose population variance σ^2 is known (and $n \geq 30$ ideally), then

$$\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

is a 95% percent confidence interval estimator of μ. In other words,

$$\mathbb{P}\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

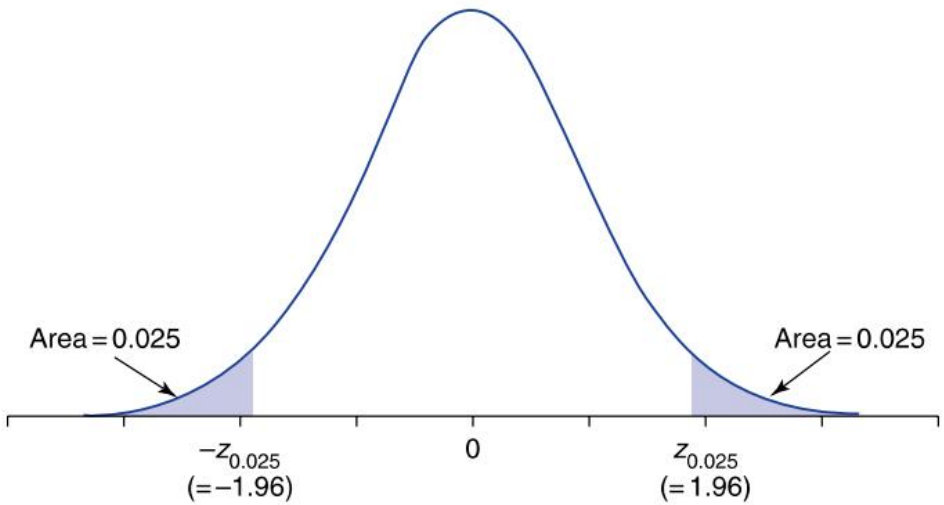


FIGURE 8.1

$P\{|Z| \leq 1.96\} = P\{-1.96 \leq Z \leq 1.96\} = 0.95.$

- 90% : 1.645 95% : 1.96 99% : 2.576

Case 1. Estimate mean μ when variance σ^2 is known

Exercise: To estimate μ, the average nicotine content of a newly marketed cigarette, 44 of these cigarettes are randomly chosen, and their nicotine contents are determined. Assume that it is known from past experience that the standard deviation of the nicotine content of a cigarette is equal to 0.7 milligrams.
If the average nicotine finding is 1.74 milligrams, what is a 95 percent confidence interval estimator of μ ?

- The length of the 95% confidence interval is $2 \times 1.96 \frac{\sigma}{\sqrt{n}}$
- If we want the length less than b, then n must be at least $\left(\frac{2 \times 1.96}{b} \right)^2$

Exercise (cont): How large a sample is necessary for the length of the 95 percent confidence interval to be less than or equal to 0.3 milligrams?

More exercises

Exercise: From past experience it is known that the weights of salmon grown at a commercial hatchery are normal with a mean that varies from season to season but with a standard deviation that remains fixed at 0.3 pounds. If we want to be 90 percent certain that our estimate of the mean weight of a salmon is correct to within ± 0.1 pounds, how large a sample is needed? What if we want to be 99 percent certain?

Exercise: The standard deviation of the lifetime of a certain type of light bulb is known to equal 100 hours. A sample of 169 such bulbs had an average life of 1350 hours. Find a 90 percent confidence interval estimate of the mean life of this type of bulb.

Exercise: A pilot study has revealed that the standard deviation of workers' monthly earnings in the chemical industry is \$180. How large a sample must be chosen to obtain an estimator of the mean salary that, with 90 percent confidence, will be correct to within $\pm$\$20?

Special case: estimating population proportion

Recall sampling proportion: when people chose between two options A or B.

- Population proportion is p
- $\overline{X}$ has expected value p and variance $\dfrac{p(1-p)}{n}$

A 95% confidence interval estimator of p is

$$\left(\overline{X} - 1.96\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}, \overline{X} + 1.96\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right)$$

Advantage: In this case, we don't need to know population variance σ.

Exercise: Out of a random sample of 100 students at a university, 82 stated that they were nonsmokers. Based on this, construct a 99 percent confidence interval estimate of p, the proportion of all the students at the university who are nonsmokers.

More exercises

Exercise: A wine importer has the opportunity to purchase a large consignment of 1947 Chateau Lafite Rothschild wine. Because of the wine's age, some of the bottles may have turned to vinegar. However, the only way to determine whether a bottle is still good is to open it and drink some. As a result, the importer has arranged with the seller to randomly select and open 20 bottles. Suppose 3 of these bottles are spoiled. Construct a 95 percent confidence interval estimate of the proportion of the entire consignment that is spoiled.

The number of necessary samples to ensure the length of the 95% interval at most b is

$$n > (\frac{1.96}{b})^2$$

Exercise: What is the smallest number of death certificates we must randomly sample to estimate the proportion of the U.S. population that dies of cancer, if we want the estimate to be correct to within 0.01 with 95 percent confidence

Case 2. Estimate mean μ when variance σ^2 is unknown

- If variance σ^2 of the true population is unknown, compute sample variance $s^2 = \dfrac{\sum(X_i - \overline{X})^2}{n-1}$ instead.
- Set $\alpha = 1 - \frac{95}{100} = 0.05$ for 95% (similary $\alpha = 0.1$ for 90%, etc)
- Look up value $t_{n-1,\alpha}$ on Student's t table: http://www.sthda.com/french/wiki/table-de-student-ou-table-t
- The confident level is

$$\left(\overline{X} - t_{n-1,\alpha}\frac{s}{\sqrt{n}}, \overline{X} + t_{n-1,\alpha}\frac{s}{\sqrt{n}}\right)$$

Exercise: The National Center for Educational Statistics recently chose a random sample of 2000 newly graduated college students and queried each one about the time it took to complete his or her degree. If the sample mean was 5.2 years with a sample standard deviation of 1.2 years, construct a 95 percent confidence interval estimate of the mean completion time of all newly graduated students

More exercises

Exercise: The manager of a shipping department of a mail-order operation located in New York has been receiving complaints about the length of time it takes for customers in California to receive their orders. To learn more about this potential problem, the manager chose a random sample of 12 orders and then checked to see how many days it took to receive each of these orders. The resulting data were

$$15, 20, 10, 11, 7, 12, 9, 14, 12, 8, 13, 16$$

Find a 90 percent confidence interval estimate for the mean time it

Exercise: A large company self-insures its large fleet of cars against collisions. To determine its mean repair cost per collision, it has randomly chosen a sample of 16 accidents. If the average repair cost in these accidents is \$2200 with a sample standard deviation of \$800, find a 90 percent confidence interval estimate of the mean cost per collision.

Exercise: To determine the average time span of a phone call made during mid-day, the telephone company has randomly selected a sample of 1200 such calls. The sample mean of these calls is 4.7 minutes, and the sample standard deviation is 2.2 minutes. Find a 95 percent confidence interval estimate of the mean length of all such calls.

Test d'hypothèse

Test d'hypothèse

- A *Hypothèse* (*hypothesis*) is a statement about a population parameter (such as mean μ or variable σ^2).
- The original hypothesis is called l'*hypothèse nulle* (*null-hypothesis*) H_0. Eg. $H_0 : \mu \leq 1.5$
- L'*hypothèse alternative* (*alternative hypothesis*) H_1 est l'hypothèse complémentaire à l'hypothèse nulle. Eg. $H_1 : \mu > 1.5$
- Objective: Test to either accept null hypothesis H_0 or reject null hypothesis (i.e. accept alternative hypothesis H_1)

Exercise: A British pharmaceutical company, Glaxo Holdings, has recently developed a new drug for migraine headaches. Among the claims Glaxo made for its drug, called somatriptan, was that the mean time needed for it to enter the bloodstream is less than 10 minutes. To convince the Food and Drug Administration of the validity of this claim, Glaxo conducted an experiment on a randomly chosen set of migraine sufferers. To prove the company's claim, what should Glaxo have taken as the null and the alternative hypotheses?

Criteria: Accept H_0 95% of times

Niveau de signification (*significance-level*) α: probability H_0 is rejected. Eg. $\alpha = 0.05$ means 95% accepted H, similar to 95% confidence interval.

- Hypotheses: $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ for some value μ_0.
- Population variance σ^2 is known
- niveau de signification is set $\alpha = 0.05$
- Reject H_0 if

$$\frac{\sqrt{n}}{\sigma}\left|\overline{X} - \mu_0\right| \geq z_{\alpha/2} = 1.96$$

- Accept H_0 otherwise.

Exercise: Suppose that the intensity of received signal is is normally distributed with mean μ and standard deviation 4. It is suspected that μ is equal to 10. Test whether this hypothesis is plausible if the same signal is independently received 20 times and the average of the 20 values received is 11.6. Use the 5% niveau de signification.

More exercises

Exercise: A leasing firm operates on the assumption that the annual number of miles driven in its leased cars has mean 13,500 and standard deviation 4000 miles. To see whether this assumption is valid, a random sample of 36 one-year-old cars has been checked. What conclusion can you draw if the average mileage on these 36 cars is 15,233?

Exercise: Traffic authorities claim that traffic lights are red for a time that has mean 30 seconds and standard deviation 1.4 seconds. To test this claim, a sample of 40 traffic lights was checked. If the average time of the 40 red lights observed was 32.2 seconds, can we conclude, at the 5 percent level of significance, that the authorities are incorrect? What about at the 1 percent level of significance.

Find the right significant-level

Recall:

- Reject H_0 if
$$\frac{\sqrt{n}}{\sigma}\left|\overline{X} - \mu_0\right| \geq z_{\alpha/2}$$
- Accept H_0 otherwise.

In practice, the significance level is often not set in advance.

p-valeur (*p-value*) p = niveau de significance α where obtain borderline accept-reject:
$$\frac{\sqrt{n}}{\sigma}\left|\overline{X} - \mu_0\right| = z_{p/2}$$

Calculate probability of Z at https://www.mathsisfun.com/data/standard-normal-distribution-table.html

Exercise: Suppose that the average of the 20 values in previous example is equal to 10.8. What is p-value for the hypothesis $\mu = 10$? (Recall that $\sigma = 4$)

More exercises

Exercise: Historical data indicate that household water use tends to be normally distributed with a mean of 360 gallons and a standard deviation of 40 gallons per day. To see if this is still the situation, a random sample of 200 households was chosen. The average daily water use in these households was then seen to equal 374 gallons per day.

- Are these data consistent with the historical distribution? Use the 5 percent level of significance.
- What is the p value?

Exercise: To test the hypothesis $H_0 : \mu = 105$ against $H_1 : \mu \neq 105$ a sample of size 9 is chosen. If the sample mean is $\overline{X} = 100$, find the p-value if the population standard deviation is known to be

- $\sigma = 5$
- $\sigma = 10$

In which cases would the null hypothesis be rejected at the 5 percent level of significance?

Population proportion

- Recall: p is proportion of population favoring A over B.
- Suppose that null hypothesis $H_0 : p \leq p_0$ and $H_1 : p > p_0$ for some given value p_0.
- Then p-value $= \mathbb{P}\Big(Binom(n, p_0) \geq \overline{X}\Big)$
- Compute $Binom(n, p_0)$ distribution at https://stattrek.com/online-calculator/binomial.aspx

Exercise: A noted educator claims that over half the adult U.S. population is concerned about the lack of educational programs shown on television. To gather data about this issue, a national polling service randomly chose and questioned 920 individuals. If 478 (52 percent) of those surveyed stated that they are concerned at the lack of educational programs on television, does this prove the claim of the educator?

More exercises

Exercise: A computer chip manufacturer claims that at most 2 percent of the chips it produces are defective. An electronics company, impressed by that claim, has purchased a large quantity of chips. To determine if the manufacturer's claim is plausible, the company has decided to test a sample of 400 of these chips. If there are 13 defective chips (3.25 percent) among these 400, does this disprove (at the5 percent level of significance) the manufacturer's claim?

Exercise: A standard drug is known to be effective in 72 percent of cases in which it is used to treat a certain infection. Anew drug has been developed, and testing has found it to be effective in 42 cases out of 50. Is this strong enough evidence to prove that the new drug is more effective than the old one? Find the relevant p value.

Exercise: An economist thinks that at least 60 percent of recently arrived immigrants who have been working in the health profession in the United States for more than 1 year feel that they are underemployed with respect to their training. Suppose a random sample of size 450 indicated that 294 individuals (65.3 percent) felt they were under-employed. Is this strong enough evidence, at the 5 percent level of significance, to prove that the economist is correct? What about at the 1 percent level of significance