STP 429: Experimental Statistics

Lecture Notes

Yi Zheng, Ph.D.
Arizona State University
Mary Lou Fulton Teachers College
School of Mathematics & Statistical Sciences
Yi.Isabel.Zheng@asu.edu
2019

Lecture 1: Review of Basic Statistics

1 What is Statistics

What do you think is Statistics, the subject? (What does a Statistician do?)

Statistics is the art and science of learning from data. This involves _____, _____, _____, _____, _____, and _____ data.

There is no unique definition of Statistics. The subject itself is fairly sophisticated. Different sources will give us different definitions. Some are more comprehensive than others. Some are closer to the current practice than others. As you get more and more experience in Statistics, you will gradually form your own understanding of what Statistics is. This will always be an interesting question to think about.

2 What is Data?

The center of Statistics is Data.

Data is information we gather with experiments, surveys, and/or observations.

Experiments, surveys, and observations are three main methods of collecting data. The three methods are ordered by the amount of control we have over the data collection process.

3 Variables

Now that we've collected our data. We need to organize them as variables in order to perform the analyses.

Data contain one or more **variables**. A variable describes any characteristic that is recorded for the subjects in the study.

*The terminology variable highlights that data values **vary**.*

In statistical softwares, we typically arrange variables as columns, and rows as observations.

4 Two types of data

- **Numeric/Continuous data** are measured on a naturally occurring numerical scale.
- **Categorical data** can only be classified into one of a group of categories.

Exercise

For each of the following studies, identify (1) the method of gathering data (experiments, observations, or surveys), (2) variables of interest, and (3) the types of the variables.

1. **Engineering.** How many miles per gallon does a car model get?

2. **Politics.** Who will likely win the election?

5 Populations, samples, and random sampling

Population and sample are a pair of concepts. Understanding them is critical to statistical inferences.

A **population** is a collection of all subjects of interest.

A **sample** is a subset of the population.

A **representative sample** exhibits characteristics typical of those possessed by the population.

A **statistical inference** is an estimate, prediction, or some other generalization about a population based on information contained in a sample.

There are cases where the population is small, and descriptive analysis is sufficient. For example, if my population is the students in my class, I can simply look at the mean, the standard deviation, and the histogram of the entire population. Most of the time, though, inferential analysis is needed to generalize from the sample data to the population of interest and make a statistical conclusion, such as is the new drug more effective than the old drug, how would I predict the final election results based on the exit survey results.

6 Descriptive Statistics

- Two types of descriptive methods: numeric or graphic
- The **goal** of descriptive statistics is to reduce the data to presentable form without losing much information.
- Different types of data require different descriptive methods. (*Remember the two types of data?*)

6.1 Describing categorical data

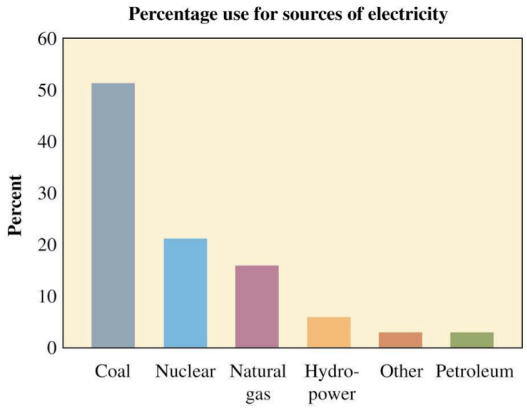
Numeric: Frequency tables

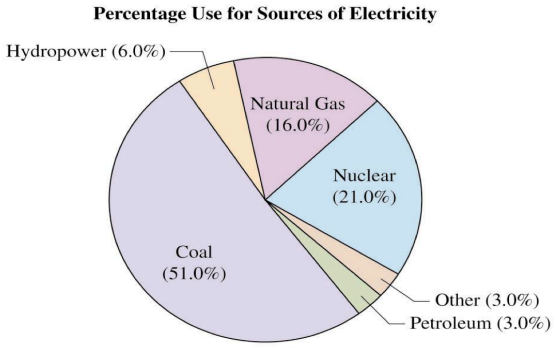
Response	Frequency	Relative Frequency	Relative Frequency (excluding missing)
Option 1	60	.60	.62
Option 2	30	.30	.31
Option 3	7	.07	.07
Missing	3	.03	
Total	100	1.00	1.00

Graphs

Bar graph presents the frequencies or percentages in each category by the height of the bars. It is helpful when we want to see the order of the frequencies of the categories or how the categories compare against each other. (*The bars are disjunct.*)

Pie charts uses the sizes of the wedges to present the percentages out of the whole. It is helpful when we want to see the portions of the categories compared to the whole.





6.2 Describing numeric data

The **Mean**

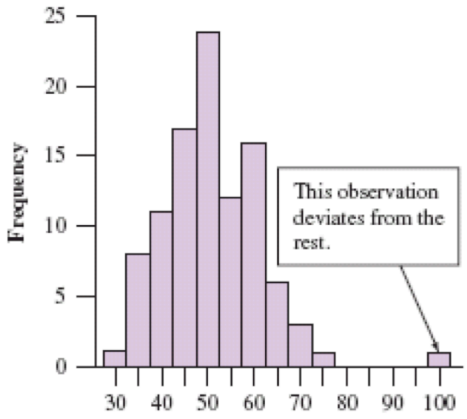
- Sample mean:
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$
- Population mean: $\mu = E(x)$
- *Note*: **Population parameter** (_____ letters) vs. **sample statistics** (_____ letters)
- The mean measures the center of the data points; It's the single most representative value.
- Other measures of center: median and mode.
  - The mean is more commonly used because it makes use of all data points.
  - The median is preferred when the data is heavily skewed.

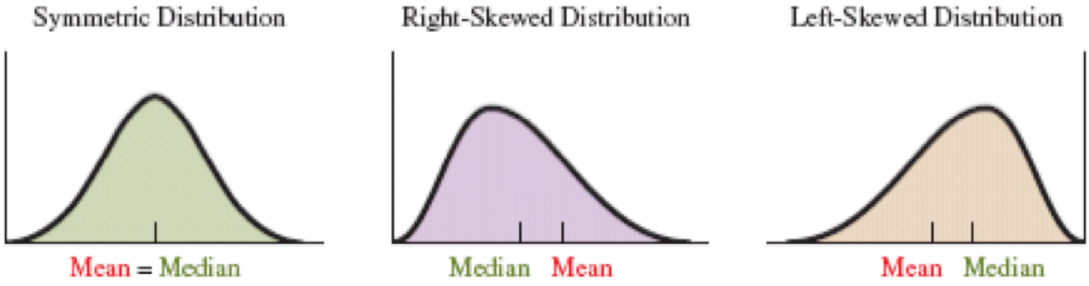
The **Standard deviation**

- Sample standard deviation:
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$
- Population standard deviation: $\sigma = \sqrt{E[(x-\mu)^2]}$
- The standard deviation represents the "average distance" of an observation from its mean. It measures the spread of data.

Graphs

Histograms present the distribution of the data. From a histogram, we can get an idea of the _____, _____, _____, _____, and _____ of the data.





6.3 Describing relationship between two numeric variables

The **Correlation coefficient** describes how closely two quantitative variables relate to each other linearly.

- Sample correlation coefficient:

$$r = \frac{\text{Cov}(x,y)}{s_x s_y} = \frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)}}{s_x s_y}$$

- Population correlation coefficient:

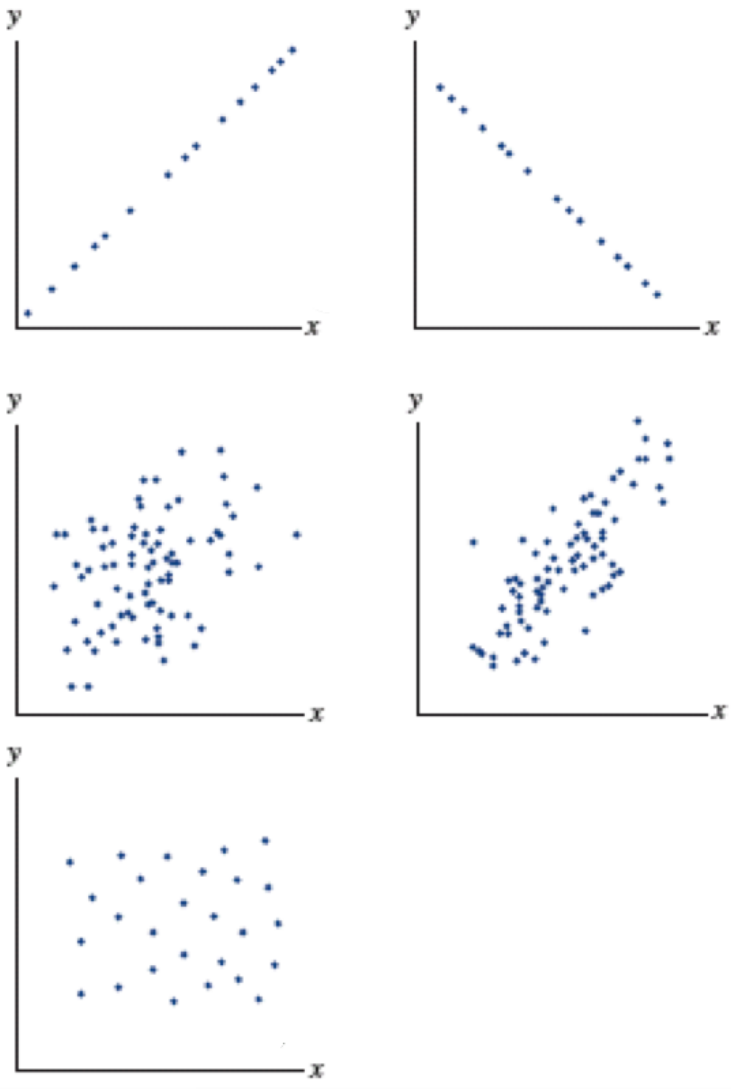
$$\rho = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sqrt{E[(x-\mu_x)^2]E[(y-\mu_y)^2]}}$$

- The sign of r indicates the direction of the correlation.
- The absolute value of r indicates the strength of the correlation.

Exercise

Match the following Pearson's correlation coefficient values with each scatter plot below.

$$r = -1, 0, 0.4, 0.8, 1$$



7 The Normal Distribution

The normal distribution is *symmetric*, *bell-shaped*, and characterized by its mean μ and standard deviation σ.

The normal distribution is the most important distribution in statistics. Many distributions have an approximate normal distribution.

The p.d.f. (probability density function) of normal distribution is

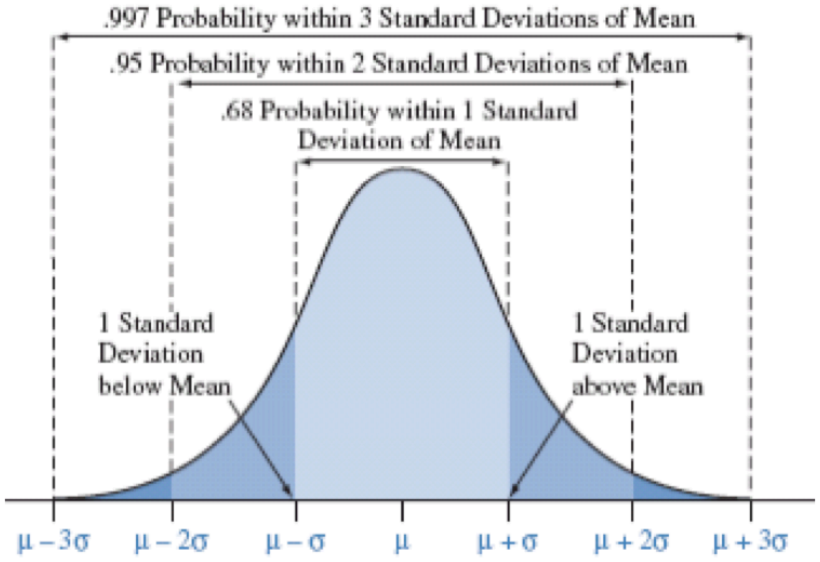
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The mean and the standard deviation completely describe the density curve.

- Increasing/decreasing μ moves the curve along the horizontal axis
- Increasing/decreasing σ controls the spread of the curve

The 1-2-3 standard deviation rule:

- 68% of the observations fall within 1 standard deviation of the mean.
- 95% fall within 2 standard deviations of the mean.
- 99.7% fall within 3 standard deviations of the mean.



Exercise

1. What is the probability that a random normal variable falls greater than its mean? $P(X > \mu)$
2. What is the probability that a random normal variable falls less than its mean minus one standard deviation? $P(X < \mu - \sigma)$
3. What is the probability that a random normal variable falls greater than its mean plus one standard deviation? $P(X > \mu + \sigma)$

---

z-distribution (a.k.a., the standard normal distribution)

For the ease of analysis, we often convert a normal variable to z**-score**. The z-score is obtained by

$$z = \frac{x - \mu}{\sigma}$$

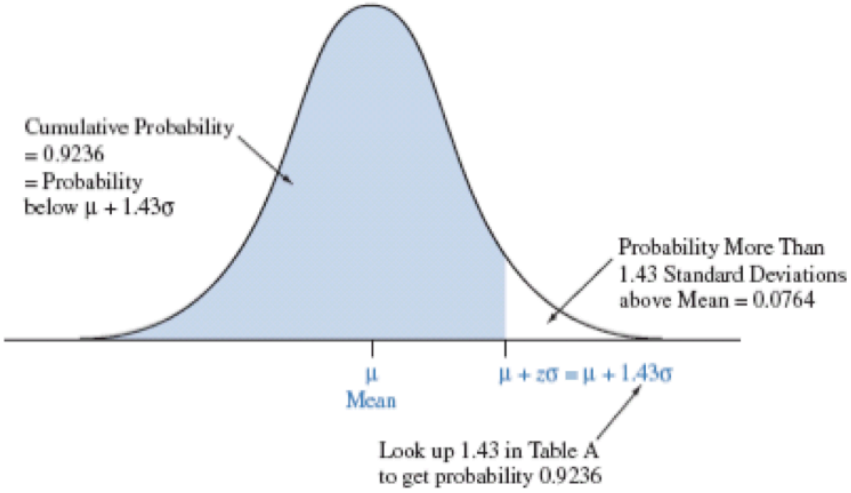
- The z value measures the distance from x to μ in terms of the number of standard deviations.
- The distribution of z has mean $\mu = 0$ and standard deviation $\sigma = 1$.
- The z-table gives the cumulative probabilities of a z-score falling below a certain value.

Example

Height of adult women can be approximated by a normal distribution. $\mu = 65$ in, $\sigma = 3.5$ in. What proportion of women are taller than 70 inches?

Solution:

Then look up the z-table.



8 Sampling Distribution

A **sample statistic** is a numerical summary of sample data. (e.g., sample mean $\bar{x}$, sample standard deviation s, sample proportion p)

A **population parameter** is a numerical summary of the population. (e.g., population mean μ, population standard deviation σ, population proportion π, regression parameters β).

In practice, we rarely know the values of the parameters. We use sample statistics to estimate population parameters. e.g., $\bar{x} \to \mu \quad s \to \sigma \quad p \to \pi$

With careful randomization and other sampling designs, we hope the one sample we draw can be as representative of the population as possible, so our sample statistic value can be as close to the true population parameter value as possible. However, unfortunately, because the sample is only a subset of the population, our sample statistic always contains random error, and the sample statistic value varies by different samples. These varied values form a sampling distribution:

The **sampling distribution** is the probability distribution of a sample statistic.

From the population, we take a sample, calculate the mean, and record the value; then we take another sample, calculate the mean, and record the value... Repeat the process for a large number of times. These values of the sample means form a distribution. That's the sampling distribution of the sample mean.

View `https://www.youtube.com/watch?v=Zbw-YvELsaM` for further illustration.

The **standard error of estimate** is the standard deviation of the sampling distribution.

9 The Central Limit Theorem

The **Central Limit Theorem**: Given a random sample of size n from a population having mean μ and standard deviation σ, as the sample size increases, the sampling distribution of the sample mean $\bar{x}$ approaches an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$, regardless of the probability distribution of the population data.

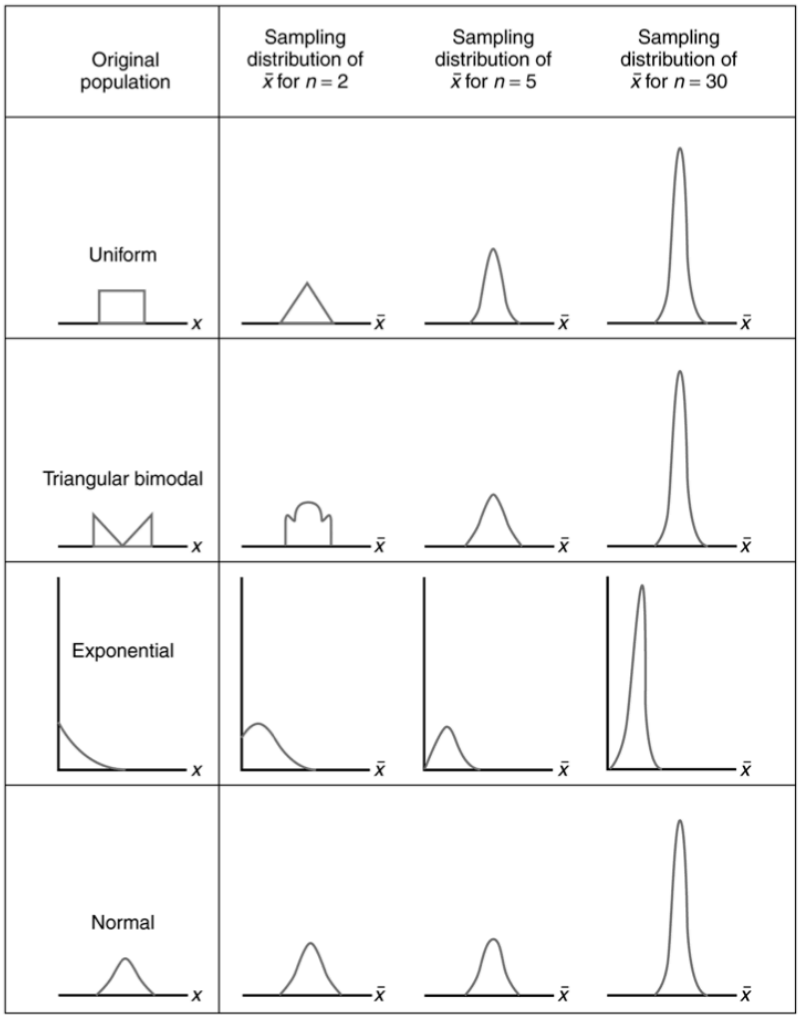
$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}} = \sigma/\sqrt{n}$$

CLT is the reason why normal distribution is the most common distribution we deal with.

CLT makes our life much easier. Regardless of the data distribution, as long as we have a large enough sample, we can resort to CLT. This is such a nice property.

How large a sample?

- The sampling distribution of the sample mean approximate normal as n _____. (Note: compare within each line in the figure below.)
- The more skewed the population distribution, the _____ n must be before the shape of the sampling distribution is close to normal. (Note: compare between the lines.)
- If the population distribution is approximately normal, then the sampling distribution is approximately normal for all sample sizes.



10 Parameter Estimation: Point Estimate and Confidence Interval

Parameter estimation and hypothesis testing are two major approaches in basic inferential statistics.

Parameter estimation further breaks down to two parts: point estimate and interval estimate (a.k.a., confidence interval).

A **point estimate** is a single number that is the "best guess" for the parameter.

- e.g., The best point estimate of the population mean μ is the sample mean $\bar{x}$.
- A point estimate doesn't tell us how close the estimate is likely to be to the parameter. Any one point estimate may or may not be close to the parameter it estimates.

A **confidence interval** is an interval containing the most believable values for a parameter.

Confidence Interval = Point Estimate $\pm$ Margin of Error
= Point Estimate $\pm$ Distribution Multiplier * Standard Error of Estimate

- An interval estimate is more useful than a point estimate. The margin of error helps us gauge the accuracy of the point estimate.

10.1 Confidence interval for a population mean

The point estimate of μ is $\bar{x}$.

Confidence interval. By CLT, for a large sample from any population OR an any-size sample from a normal underlying population, the sampling distribution of $\bar{x}$ is normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. So the $(1-2\alpha)\%$ confidence interval of μ is:

$$\text{CI}_\mu = \bar{x} \pm z_\alpha \frac{\sigma}{\sqrt{n}}$$

Note: In z_α: z is the distribution, α is the right-tail probability.

For 95% CI, $\alpha = .025$, $z_\alpha = 1.96$.

In practice, we don't know the value of σ. We use the sample standard deviation s to substitute, which introduces extra error.

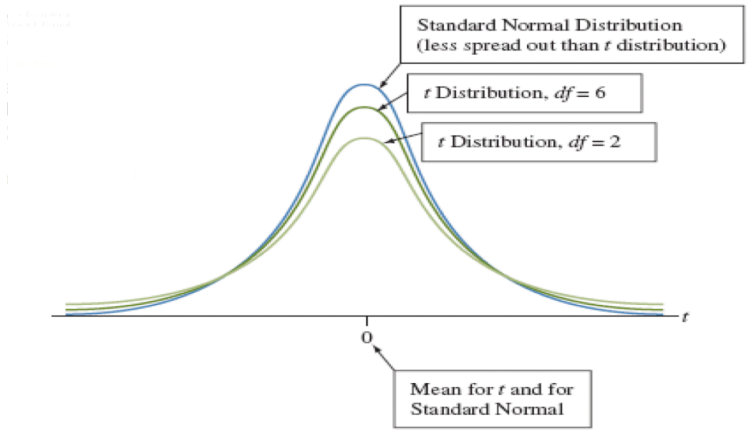
To account for this increased error, we replace the z-score the t-score. t_α is larger than z_α, so the confidence interval is a bit wider.

$$\text{CI}_\mu = \bar{x} \pm (t_{\alpha, df=n-1})(\frac{s}{\sqrt{n}})$$

---

t-distribution:

- The t-distribution is bell shaped and symmetric about 0.
- The probabilities depend on the degrees of freedom.
- The t-distribution has thicker tails than the z-distribution.



Example A study of 7 American adults yields an average height of 67.2 inches and a standard deviation of 3.9 inches. Assuming the heights are normally distributed, what is the 95% confidence interval for the average height of all American adults?

Solution: $n = 7$, $\bar{x} = 67.2$, $s = 3.9$, df$= 7 - 1 = 6$.

The 95% confidence interval for μ is:

$$\bar{x} \pm t_{.025,df=6}\frac{s}{\sqrt{n}} = 67.2 \pm 2.447(\frac{3.9}{\sqrt{7}}) = 67.2 \pm 3.6 = (63.6, 70.8)$$

Interpretation: We are 95% confident that the average height of all American adults is between 63.6 and 70.8 inches.

What if we want to be 99% confident?

The 99% confidence interval for μ is:

$$\bar{x} \pm t_{.005,df=6}\frac{s}{\sqrt{n}} = 67.2 \pm 3.707(\frac{3.9}{\sqrt{7}}) = 67.2 \pm 5.5 = (61.7, 72.7)$$

Note: With the same sample size, if we want to increase our confidence level, we will have to widen the interval. Like basketball shooting, larger basket, more confidence.

What if our sample size is 50?

The 95% confidence interval for μ is:

$$\bar{x} \pm t_{.025,df=49}\frac{s}{\sqrt{n}} = 67.2 \pm 2(\frac{3.9}{\sqrt{50}}) = 67.2 \pm 1.1 = (66.1, 68.3)$$

Note: Increasing the sample size and increase the precision of our estimate.

10.2 Confidence interval for a population proportion

Point estimate of population proportion π is the sample proportion p.

Standard error of p is $\sqrt{\frac{\pi(1-\pi)}{n}}$.

In practice, we don't know π, so we use p to substitute.

$$\text{CI}_\pi = p \pm z_\alpha\sqrt{\frac{p(1-p)}{n}}$$

11 Hypothesis Testing

Five steps of hypothesis testing:

1. Check the assumptions
2. Construct the hypotheses
3. Calculate the test statistic
4. Find the p-value
5. Make the conclusion

11.1 Hypothesis testing of a population mean

Step 1: Assumptions

Note: Content in boxes are examples with a one-sample t-test.

- The data are obtained using randomization
- The sample size is large ($n \geq 30$)

Step 2: Hypotheses

- A hypothesis is a statement about a population, usually of the form that a certain parameter takes a particular numerical value or falls in a certain range of values.
- The null hypothesis (H_0) is a statement that the parameter takes a particular value. It has a single parameter value.
- The alternative hypothesis (H_1) states that the parameter falls in some alternative range of values.

H_0: $\mu = \mu_0$

H_1: $\mu > \mu_0$ or $\mu < \mu_0$ or $\mu \neq \mu_0$

Step 3: Test Statistic

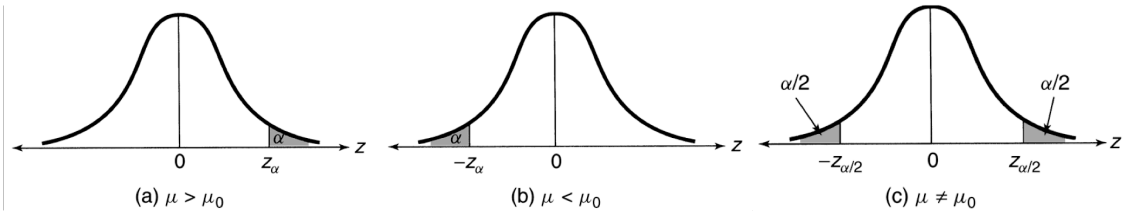
- A test statistic describes how far the point estimate falls from the hypothesized parameter value if the null hypothesis is true (usually in terms of the number of standard errors between the two).
- If the point estimate falls far from the value suggested by the null hypothesis in the direction specified by the alternative hypothesis, it is good evidence against the null hypothesis and in favor of the alternative hypothesis.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad df = n - 1$$

Step 4: p-value

- The p-value is the probability that the test statistic equals the observed value or a value even more extreme.
- A small p-value indicates that the observed test-statistic is not a plausible outcome from the hypothesized sampling distribution.
- The smaller the p-value, the less likely the observed data are if H_0 were true, and the strong the evidence is against H_0.

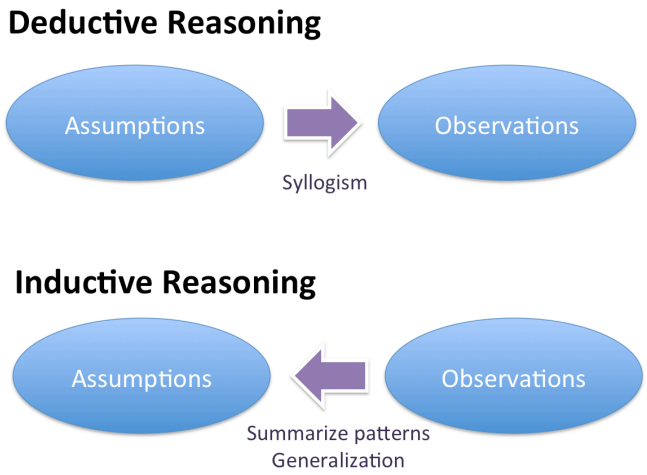
H_1: $\mu > \mu_0$	Right-tail probability from the t-distribution.
H_1: $\mu < \mu_0$	Left-tail probability from the t-distribution.
H_1: $\mu \neq \mu_0$	Two-tail probability from the t-distribution.



Step 5: Conclusion

- If the p-value is smaller than a preset threshold (typically 0.05), reject the null hypothesis (H_0) and conclude that there is an effect as described by the alternative hypothesis (H_1).
- If the p-value is greater than a preset threshold (typically 0.05), we fail to reject the null hypothesis (H_0) and conclude that there is not enough evidence for such effect as described by the alternative hypothesis (H_1).

11.2 Logic of hypothesis testing



Deductive reasoning

- Deductive reasoning is the dominant form in mathematics. e.g.,
  - x + 5 = 8 is a kind of assumption.
  - We solve for x = 3 based on the assumption.
  - The solution is correct as long as the assumption is correct.
- Syllogism
  - All men are mortal.
  - David is a man.
  - David is mortal.
- Deductive reasoning is rigorous and strong.

Inductive reasoning

- Most knowledge of the world comes form induction.
- But inductive conclusions are not necessarily correct:

  You can go to Australia and see that kangaroos hop on two legs. Every kangaroo you see is hopping on two legs. You conclude, inductively, that all kangaroos hop on two legs. There might be one-legged kangaroos. That you havent seen them doesnt mean they cant exist.

In hypothesis testing:

We want to make inference of the assumption based on our observed data — We want to go from observations to assumptions.

But this is an inductive reasoning and it is not always correct!

So we want to resort to deductive reasoning for a more rigorous reasoning. However, deductive reasoning goes from assumption to observation!

Contrapositive

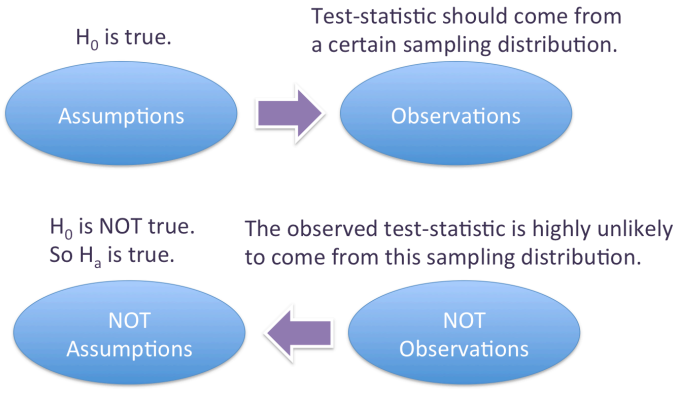
Contrapositive is a way of recasting a statement in a new form that will be true so long as the original statement is true.

Statement: If it is my car, then it is red.
Contrapositive: If it is NOT red, then it is NOT my car.

Simple rule: negate both parts and then switch the order.

Other forms of recast are are not necessarily true:

- If it is red, then it is my car.
- If it is NOT my car, then it is NOT red.



Note

- If not enough evidence is obtained for rejecting H_0, we cannot reasonably accept H_0. (If the car is red, it is mine.)
- The only valid conclusion of a hypothesis testing is rejecting H_0. So make sure to pick an H_0 hypothesis that will be meaningful to reject.
- There is not a completely clean way to say whether the observations fail to match the consequences of the null hypothesis.

11.3 Comparing two group means

Example A researcher was interested in comparing the resting heart rate of people who exercise regularly and people who do not exercise regularly. Independent simple random samples of 16 people ages 30-40 who do not exercise regularly and 12 people ages 30-40 who do exercise regularly were selected and the resting heart rate of each person was measured. The summary statistics are as follows.

Group	Mean	Std	n
Do exercise	69.3	8.7	12
Do not exercise	73.5	10.2	16

Is the mean resting pulse rate of people who exercise regularly lower than the mean resting pulse rate of people who do not exercise regularly?

Step 1: Assumptions

- Independent random samples
- Independent groups
- Approximately normal population distributions for each group

Step 2: Hypotheses

H_0: $\mu_Y = \mu_N$

H_1: $\mu_Y \neq \mu_N$

Step 3: Test Statistic

$$t = \frac{\bar{x}_Y - \bar{x}_N}{\sqrt{\frac{s_Y^2}{n_Y} + \frac{s_N^2}{n_N}}} = \frac{69.3 - 73.5}{\sqrt{\frac{8.7^2}{12} + \frac{10.2^2}{16}}} = -1.17$$

df = the smaller of $n_Y - 1$ and $n_N - 1$ = 11

Step 4: p-value

Because H_1: $\mu_Y \neq \mu_N$, the p-value is the two-tail probability.

In the t-distribution with df=11, the two-tail probability of -1.17 is 0.26.

Step 5: Conclusion

With a significance level of 0.05, because our p-value $0.26 > 0.05$, we fail to reject H_0.

We conclude that there is not enough evidence for the claim that the mean resting pulse rate of people who exercise regularly lower than the mean resting pulse rate of people who do not exercise regularly.

Question: Do you still believe regular exercise would give us a stronger heart? How can you improve the study to provide enough statistical evidence for your claim?

11.4 Statistical significance vs. practical significance

- The significance test gives us information about whether the parameter differs from the H_0 value, but it does not tell us about the practical importance of the results.
- When the sample size is very large, tiny deviations from the null hypothesis (with little practical consequence) may be found to be statistically significant.
- When the sample size is very small, large deviations from the null hypothesis (of great practical importance) might go undetected (statistically insignificant).

Effect size, a measure of practical significance, is a useful index to accompany hypothesis testing. For example,

$$\text{Cohen's d} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}}$$

For more information on effect sizes, see `https://en.wikipedia.org/wiki/Effect_size`