Preliminaries: Inferential Statistics

CS 418. Introduction to Data Science

Preliminaries
Descriptive vs. Inferential Statistics

- **Statistics** deals with the collection, analysis, interpretation, and presentation of data.
- There are two types of statistical analysis: **descriptive statistics** and **inferential statistics.**
- **Descriptive statistics** focuses on **summarizing data** about a sample drawn from a population.
- **Inferential statistics** focuses on using information from the sample to **make conclusions** about the population from which the sample was drawn.
  - One of the primary methods of **inferential statistics** is **hypothesis testing**, which allows us to compare differences between population parameters.
    - A **parameter** is a numerical characteristic of a population (e.g., **mean**, **median**, **standard deviation**).

Preliminaries
Hypothesis Testing (I)

- A **hypothesis** is a statement that makes a claim about the parameters of one or more populations.
- To determine whether a hypothesis should be retained or rejected, we perform a **hypothesis test**, which compares two competing hypothesis:
  - The **null hypothesis**, denoted $\boldsymbol{H_0}$, is a statement assumed to be true unless sufficient evidence indicates otherwise.
    - Typically, the **null hypothesis** asserts that the true value of a population parameter is **equal** to a hypothesized value or that the parameters for two populations are **not different**.
  - The **alternative hypothesis**, denoted $\boldsymbol{H_a}$, is a statement that contradicts H_0.
    - Typically, the **alternative hypothesis** asserts that the true value of a population parameter is **not equal** to a hypothesized value or that the parameters for two populations are **different**.

Preliminaries
Hypothesis Testing (II)

- If there is sufficient evidence favoring the **alternative hypothesis**, we **reject the null hypothesis**.
- Otherwise, we **fail to reject the null hypothesis**.
- ***Example*:**
  - We want to test whether the mean GPA of college students is different from 2.0. The **null hypothesis** is $\boldsymbol{H_0: \mu = 2.0}$ and the **alternative hypotheses** is $\boldsymbol{H_a: \mu \neq 2.0}$.
  - We want to test whether college students take less that five years to graduate on average. **What are the null and alternative hypotheses?**

    $\boldsymbol{H_0}$: $\mu = 5$

    $\boldsymbol{H_a}$: $\mu < 5$

Preliminaries
Hypothesis Testing (III)

- The basic steps of performing a **hypothesis test** are:
  1. Specify the **null hypothesis** H_0 and the **alternative hypothesis** H_a.
  2. Specify the **significance level** α.
  3. Collect the **data**.
  4. Calculate the **test statistic**.
  5. Calculate the corresponding p**-value**.
  6. Compare the p**-value** with the **significance level**.
  7. Determine whether to **reject** or **fail to reject** the **null hypothesis** H_0.

Hypothesis Testing
Type I and Type II Errors (I)

- A **hypothesis test** has four possible outcomes:

	H_0 is true	H_0 is false
Reject H_0	**Type I Error**	Correct Decision
Fail to reject H_0	Correct Decision	**Type II Error**

- Each of the errors occurs with a particular probability:
  - α is the probability of rejecting the null hypothesis when the null hypothesis is true (**type I error**).
  - β is the probability of failing to reject the null hypothesis when the null hypothesis is false (**type II error**).
- The **significance level** of a hypothesis test is α.
- The **power** of a hypothesis test is $1 - \beta$.

Hypothesis Testing
Type I and Type II Errors (II)

- ***Example*:**
  - Suppose that the **null hypothesis** is that your rock climbing equipment is safe. The **type I error** is thinking that the rock climbing equipment is not safe when, in fact, it is safe. The **type II error** is thinking that the rock climbing equipment is safe when, in fact, it is not safe. In this case, the error with the greater consequence is the **type II error**.
  - Suppose that the **null hypothesis** is that a patient is sick. **What are the type I and type II errors?**

    Type I error: THINKING THAT A SICK PATIENT IS HEALTHY

    Type II error: THINKING THAT A HEALTHY PATIENT IS SICK

    Which error has the greater consequence? TYPE I ERROR

Preliminaries
P-Value (I)

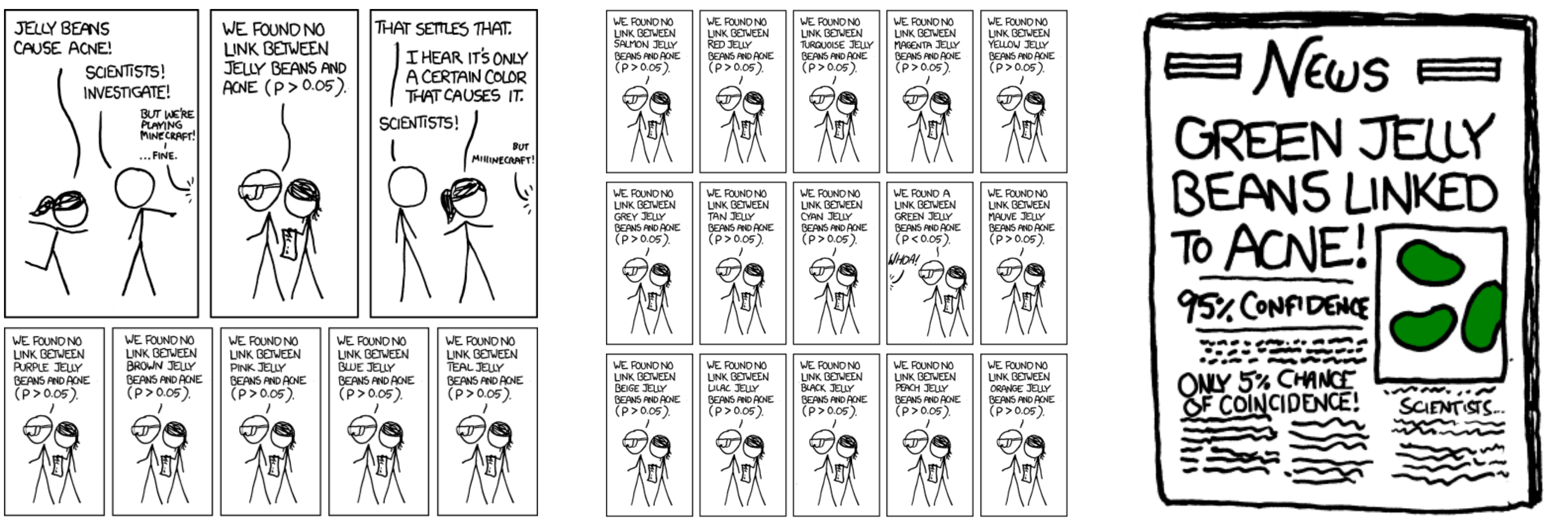
- A ***p*-value** is the probability of obtaining a result that is as extreme or more extreme than the observed data if the null hypothesis were true.
- If the ***p*-value** is less than the specified **significance level**, then two possibilities exist:
  - The **null hypothesis is true** and the observed data is relatively unusual simply due to chance.
  - The **null hypothesis is false** and the alternative hypothesis provides a more reasonable explanation of the observed data.
- A **statistically significant** result differs enough from the null hypothesis that a conclusion can be inferred about the population.
- **Statistical significance** does not necessarily imply **practical significance** of a result.

Preliminaries
P-Value (II)

Beware of the danger zone!

"If you're setting out to find 'significant' results, you usually can."

- Testing many hypotheses until finding one that is statistically significant is a common misuse of ***p*-values** known as ***p*-hacking**.
- You should determine your hypotheses and significance level **before looking at the data**.



Source: xkcd, *Significant*

Preliminaries
P-Value (III)

Beware of the danger zone!

"p-values are not substitutes for common sense."

- The American Statistical Association (ASA) provides a set of principles for the proper use and interpretation of **p-values**, including:
  - "**p-values** do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone."
  - "Scientific conclusions and business or policy decisions should not be based only on whether a **p-value** passes a specific threshold."
  - "Proper inference requires full reporting and transparency."
  - "A **p-value**, or statistical significance, does not measure the size of an effect or the importance of a result."
  - "By itself, a **p-value** does not provide a good measure of evidence regarding a model or hypothesis."

Hypothesis Testing
z-Test for Population Means (I)

- A ***z*-test** is a hypothesis test in which the test statistic follows a **normal distribution**.
- A ***z*-test** can be used to determine whether the **population mean μ is equal to a hypothesized mean μ_0**, assuming that the **population standard deviation σ is known**
- To perform a ***z*-test** , the following **assumptions** should be met:
  - The **population standard deviation σ is known**.
  - **Randomness**. Data comes from a simple, random sample.
  - **Independence**. The value of each observation does not affect the value of other observations
  - **Normality**. Population is approximately normally distributed or the sample size is sufficiently large (more than 30 observations).

Hypothesis Testing
z-Test for Population Means (II)

- Set the **null and alternative hypotheses**:

  $H_0: \mu = \mu_0$

  $H_a: \mu_1 \neq \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$

  where μ is the population mean and μ_0 is the hypothesized population mean.

- Compute the **z-test statistic**:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

σ is the **population** standard deviation

- Compute the corresponding **p-value**.
- Make a decision given a previously selected **significance level α**.
  - If p**-value** $< \alpha$, **reject** the null hypothesis.
  - If p**-value** $\geq \alpha$, there is **no sufficient evidence to reject** the null hypothesis.

z-Test for Population Means
Example (I)

- ***Example***: Suppose that you want to determine whether a smartphone has less than the **7.8-hour** battery life claimed by the manufacturer. You sampled **10 smartphones** with a mean battery life of **7.7 hours**. The population standard deviation of the battery life is **0.57 hours**. **Does sufficient evidence exist that the battery life of the smartphone is lower than the manufacturer's claim at a significance level of $\alpha = 0.05$?**
  - Set the **null and alternative hypotheses**:

    H_0: $\mu = 7.8$

    H_a: $\mu < 7.8$

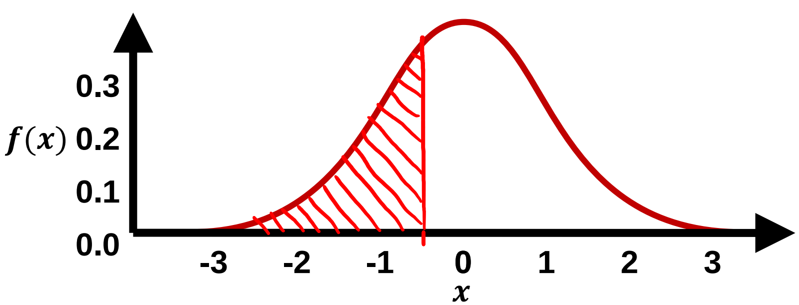
    Left-tailed test
  - Compute the **z-test statistic**:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{7.7 - 7.8}{0.57/\sqrt{10}} \approx -0.555$$

z-Test for Population Means Example (II)

- ***Example***: Suppose that you want to determine whether a smartphone has less than the **7.8-hour** battery life claimed by the manufacturer. You sampled **10 smartphones** with a mean battery life of **7.7 hours**. The population standard deviation of the battery life is **0.57 hours**. **Does sufficient evidence exist that the battery life of the smartphone is lower than the manufacturer's claim at a significance level of $\alpha = 0.05$?**
  - Compute the corresponding **p-value**:

$$p\text{-value} = P(z \leq -0.555) \approx 0.289$$



In Python (with **scipy**):

```
import scipy.stats as st
st.norm.cdf(-0.555)
```

Click for documentation

z-Test for Population Means
Example (III)

- ***Example***: Suppose that you want to determine whether a smartphone has less than the **7.8-hour** battery life claimed by the manufacturer. You sampled **10 smartphones** with a mean battery life of **7.7 hours**. The population standard deviation of the battery life is **0.57 hours**. **Does sufficient evidence exist that the battery life of the smartphone is lower than the manufacturer's claim at a significance level of $\boldsymbol{\alpha = 0.05}$?**
  - Make a decision given **significance level $\boldsymbol{\alpha}$**.

$0.289 > 0.05$

Since the ***p*-value is greater than the significance level $\boldsymbol{\alpha = 0.05}$**, there is **no sufficient evidence to reject** the manufacturer's claim.

Hypothesis Testing
t-Test for Population Means (I)

- The **population standard deviation σ** is rarely known.
- When the **population standard deviation σ is unknown**, a **t-test** can be used to determine whether the **population mean μ is equal to a hypothesized mean μ_0**.
- A **t-test** is a hypothesis test in which the test statistic follows a **Student's t-distribution**.
- To perform a **t-test** , the following **assumptions** should be met:
  - **Randomness**. Data comes from a simple, random sample.
  - **Independence**. The value of each observation does not affect the value of other observations
  - **Normality**. Population is approximately normally distributed or the sample size is sufficiently large (more than 30 observations or less if the data is not skewed and no outliers are present).

Hypothesis Testing
t-Test for Population Means (II)

- Set the **null and alternative hypotheses**:

  $H_0: \mu = \mu_0$

  $H_a: \mu_1 \neq \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$

  where μ is the population mean and μ_0 is the hypothesized population mean.

- Compute the **t-test statistic**:

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

s is the **sample** standard deviation

- Compute the **degrees of freedom**: $df = n - 1$.
- Compute the corresponding **p-value**.
- Make a decision given a previously selected **significance level α**.
  - If **p-value** $< \alpha$, **reject** the null hypothesis.
  - If **p-value** $\geq \alpha$, there is **no sufficient evidence to reject** the null hypothesis.

Hypothesis Testing
Degrees of Freedom

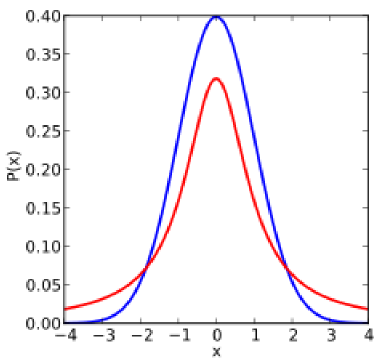
- The **degrees of freedom** refers to the number of **independent** (“**free to vary**”) values in the estimation of a parameter.
  - ***Example*:**

    Values: 1, 5, 4, 2, ***X***

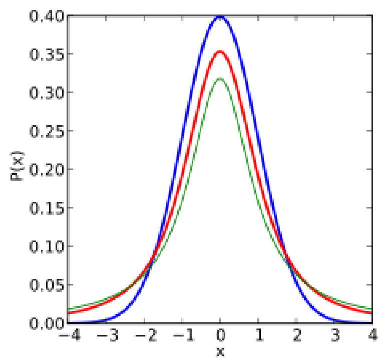
    Mean: 3

    Is *X* free to vary? No
  - Typically, the **degrees of freedom** is equal to the **sample size minus the number of intermediate parameters**.
- The shape of the ***t* distribution** is determined by its **degrees of freedom**.

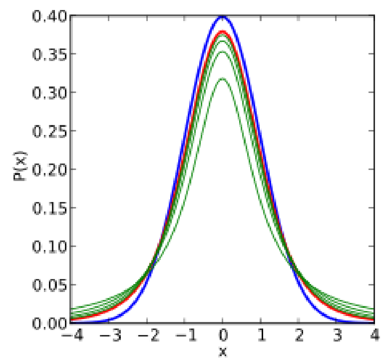
1 degree of freedom



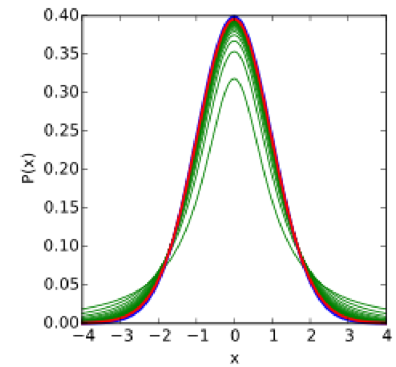
2 degrees of freedom



5 degrees of freedom



30 degrees of freedom



Preliminaries
Exercise 4.1 (I)

THINK PAIR SHARE

The mean circumference of basketballs produced in a manufacturing facility is supposed to be 29 inches. A random sample of 25 basketballs has a mean of 29.1 inches with a sample standard deviation of 0.217 inches. The quality control supervisor claims that the mean circumference of the basketballs produced in the facility is different from 29 inches.

At the $\alpha = 0.01$ significance level, does sufficient evidence exist to support the supervisor's claim?

- Set the **null and alternative hypotheses**:

H_0: $\mu = 29$

H_a: $\mu \neq 29$

- Compute the ***t*-test statistic**:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{29.1 - 29}{0.217/\sqrt{25}} = 2.304$$

- Compute the **degrees of freedom**:

$$df = n - 1 = 24$$

Preliminaries
Exercise 4.1 (II)

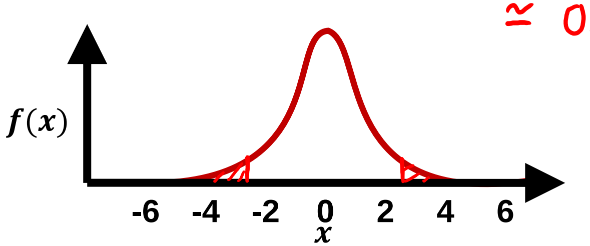
THINK PAIR SHARE

The mean circumference of basketballs produced in a manufacturing facility is supposed to be <u>29 inches</u>. A random sample of <u>25 basketballs</u> has a mean of <u>29.1 inches</u> with a sample standard deviation of <u>0.217 inches</u>. The quality control supervisor claims that the mean circumference of the basketballs produced in the facility is <u>different from 29 inches</u>.

At the $\alpha = 0.01$ significance level, does sufficient evidence exist to support the supervisor's claim?

- Compute the corresponding **p-value**:

p-value $=$ P(t < -2.304 OR t > 2.304)
$\simeq 0.0302$



In Python (with **scipy**):

```
import scipy.stats as st
st.t.cdf(x, df)
```

Click for documentation

- Make a decision given the **significance level α**: 0.0302 > 0.01

FAIL TO REJECT THE NULL HYPOTHESIS

Two-sample Hypothesis Testing

- A **one-sample hypothesis test** determines whether an observed value differs from a hypothesized value for a population.
- A **two-sample hypothesis test** determines whether there are differences between values of two populations.
- ***Example***: you randomly show visitors to a website one of two versions of an advertisement and track how many people click on each one. The goal is to determine which advertisement is more effective. This experiment setting is called **A/B testing**.
- The **test statistic** and the **standard error** for a **two-sample hypothesis test** are given by:

$$\textbf{test statistic} = \frac{\textbf{observed difference} - \textbf{hypothesized difference}}{\boldsymbol{SE}\ \textbf{for the difference}}$$

$$\boldsymbol{SE}\ \textbf{for the difference} = \sqrt{(\boldsymbol{SE_1})^2 + (\boldsymbol{SE_2})^2}$$

Two-sample Hypothesis Testing
Two-sample z-Test for Population Means

- Set the **null and alternative hypotheses**:

  $H_0: \mu_1 = \mu_2$

  $H_a: \mu_1 \neq \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$

  where μ_1 and μ_2 are means from distinct populations.
- Compute the **z-test statistic**:

$$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$$

σ_1 and σ_2 are the standard deviations of the **populations**

- Compute the corresponding **p-value**.
- Make a decision given a previously selected **significance level α**.
  - If **p-value** $< \alpha$, **reject** the null hypothesis.
  - If **p-value** $\geq \alpha$, there is **no sufficient evidence to reject** the null hypothesis.

Two-sample Hypothesis Testing

Two-sample *t*-Test for Population Means

- When the **standard deviations of the populations are unknown**, a ***t*-test** can be used to compare the means of the two populations.
- There are two types of **two-sample *t*-test**:
  - In an **unpaired *t*-test** (or **independent *t*-test**), a sample taken from one population is compared to a different sample taken from another population and exposed to the same treatment.
    - ***Example*:** a study of the difficulty of a game comparing the error rates between adults and children.
  - In a **paired *t*-test** (or **dependent *t*-test**), a sample taken from one population is exposed to two different treatments.
    - ***Example*:** a study of track and field athletes comparing resting heart rate with hearth rate after running a race.

Two-sample Hypothesis Testing
Unpaired t-Test for Population Means

- Set the **null and alternative hypotheses**:

  $H_0: \mu_1 = \mu_2$

  $H_a: \mu_1 \neq \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$

  where μ_1 and μ_2 are means from distinct populations.
- Compute the **t-test statistic**:

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

s_1 and s_2 are the standard deviations of the **samples**

- Compute the **degrees of freedom**: $df = n_1 + n_2 - 2$.
- Compute the corresponding **p-value**.
- Make a decision given a previously selected **significance level α**.

Preliminaries
Exercise 4.2

THINK PAIR SHARE

You measure the number of errors made while completing a memory-related task by a group of 10 people not taking a memory enhancement drug and by another group of 10 people taking the drug (see *data_lecture_04_b*). You claim that the drug is effective in reducing the number of errors in memory-related tasks.

At the $\alpha = 0.01$ significance level, does sufficient evidence exist to support your claim?

- Set the **null and alternative hypotheses**:

H_0: $\mu_1 = \mu_2$ μ_1: NO DRUG

H_a: $\mu_1 > \mu_2$ μ_2: DRUG

- Perform a **paired** or **unpaired** ***t*-test?**

UNPAIRED

- Make a decision: $0.009 < 0.01$

REJECT THE NULL HYPOTHESIS

In Python (with **scipy**):

```
import scipy.stats as st
st.ttest_ind(a, b,
equal_var)
```

Click for documentation

Preliminaries
References

- David Diez, Christopher Barr, and Mine Çetinkaya-Rundel. *OpenIntro Statistics* (2015).
- Joel Grus. *Data Science from Scratch* (2015).
- OpenStax. *Introductory Statistics* (2016).
- Ronald Wasserstein and Nicole Lazar. *The ASA's Statement on p-Values: Context, Process, and Purpose* (2016).