

Lecture 1 — Overview

Mohammad Reza Karimi

1 Overview

Statistics is about **analyzing** and drawing **conclusions** from **data**, and **quantifying the uncertainty** in those conclusions.

Types of analysis:

1. Describe data: we can summarize data with numbers (e.g., compute the mean, range, standard deviation) or make visual representations (e.g., histograms, box plots)
2. Estimation: the goal is to use limited samples to learn something about a broader population. Unlike descriptive statistics, which summarize the data you already have, estimation is concerned with extrapolation beyond the observed data. For example, instead of surveying every potential buyer of a product, we use a sample to estimate how much the entire population might be willing to pay.
3. Confidence intervals: to quantify uncertainty. Fewer samples → more uncertain, larger sample size → less uncertain.
4. Answer yes/no questions (hypothesis testing): for example, is a drug more effective than the placebo? **And what is the probability we made the wrong call?**
5. Make predictions (regression) of one variable given another: e.g., SAT score → GPA, or temperature during a heatwave → demand on the power grid. **In addition to outputting a single number, can we specify a probable range of values (confidence interval)?**
6. Classification: is this image a cat or a dog? A benign or cancerous tumor? **And what is the probability we misclassified?**
7. Causal inference: “Does X cause Y ?”
8. Survival analysis: predict some “time to an event.” For example, predicting survival time under a particular treatment, such as how long a cancer therapy

extends patients' lives. This requires enrolling participants in long-term studies that may last months or years. A key challenge is that not all participants remain in the study until the end. Rather than discarding these incomplete observations, we can account for the fact that such individuals indeed survived up to a certain point.

9. Data visualization: for understanding the data visually. This usually amounts to finding a smart way to project your points into a lower dimensional space.

Quantifying uncertainty in our conclusions is one of the main challenges in statistics!

2 The statistics pipeline

What is the data?

Independent, identically distributed (i.i.d.) samples $X_1, X_2, \dots, X_n \sim \mathbb{P}$, where \mathbb{P} is the unknown probability distribution! For example: X_i could be the effect of the drug on patient i .

What is the conclusion?

Some information about \mathbb{P} , such as the mean μ . In the drug vs placebo example, knowing $\mu > 0$ would tell us the drug has a positive effect.

Schematically:

$$\begin{array}{ccc} \text{i.i.d. data} & \longrightarrow & \boxed{\text{statistical}} \\ X_i \sim \mathbb{P} & & \boxed{\text{method}} \end{array} \longrightarrow \hat{\mathbb{P}} \approx \mathbb{P} \quad (1)$$

The fields of statistics and probability are complementary to each other in the following sense:

1. **Probability: given \mathbb{P} , what can we say about data from \mathbb{P} ?**

Example: $\mathbb{P} = \mathcal{N}(0, 1)$. Using probability, we can say a sample $X \sim \mathbb{P}$ lies in the interval $(-3, 3)$ with probability 0.997.

2. **Statistics: given data from \mathbb{P} , what can we say about \mathbb{P} ?**

Example: $X = 100$. Using statistics, we can say X is most likely not a sample from $\mathcal{N}(0, 1)$, i.e., $\mathbb{P} \neq \mathcal{N}(0, 1)$.

3 The power of aggregating data

In the last example, we concluded $\mathbb{P} \neq \mathcal{N}(0, 1)$ based on a single sample X . This isn't very informative, and with only a single sample, we can't say much. But by aggregating independent data, we can make more precise statements about \mathbb{P} .

Suppose we have n i.i.d. samples X_1, \dots, X_n with $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}[X_i] = \sigma^2$. (We will use the notation $\mathbb{V}[X]$ for variance of X). Suppose σ is known but μ is unknown, and our goal is to estimate μ from the data. How would we go about this? By computing the sample mean:

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i.$$

Let's compute the mean and variance of \bar{X}_n .

3.1 Mean and variance of i.i.d. averages

We'll show that

$$\mathbb{E}[\bar{X}_n] = \mu, \quad \mathbb{V}[\bar{X}_n] = \frac{\sigma^2}{n}.$$

This comes from the following calculations (make sure you understand each step):

$$\begin{aligned} \mathbb{E}[\bar{X}_n] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \mu, \\ \mathbb{V}[\bar{X}_n] &= \mathbb{V}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} \mathbb{V}\left[\sum_{i=1}^n X_i\right] = \frac{1}{n^2} \sum_{i=1}^n \mathbb{V}[X_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \end{aligned} \tag{2}$$

The equality in blue expresses the important property that for independent variables, the *sum of the variances is the variance of the sum*.

3.2 The Law of Large Numbers (LLN)

Together, $\mathbb{E}[\bar{X}_n] = \mu$ and $\mathbb{V}[\bar{X}_n] = \sigma^2/n$ tell us that the fluctuations of \bar{X}_n around μ get smaller and smaller as $n \rightarrow \infty$. This is expressed by the following law of large numbers (LLN):

$$\bar{X}_n \rightarrow \mu \quad \text{as } n \rightarrow \infty.$$

Thus, aggregating *independent* data helps us get a good estimator for the mean μ !

3.3 The Central Limit Theorem (CLT)

We know that the fluctuations of \bar{X}_n around μ are shrinking, but to do statistical inference (quantify uncertainty), we need more fine-grained information about the distribution of \bar{X}_n . This is where the *Central Limit Theorem* (CLT) comes in.

To motivate the CLT, note that

$$\mathbb{E}\left[\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)\right] = 0, \quad \mathbb{V}\left[\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)\right] = 1.$$

(Make sure you can do this calculation). It turns out that this scaled, centered random variable $(\sqrt{n}/\sigma)(\bar{X}_n - \mu)$ converges to our favorite distribution which also has mean 0 and variance 1: the normal distribution $\mathcal{N}(0, 1)$.

Theorem 3.1: Central Limit Theorem

Let $X_i, i = 1, \dots, n$ be i.i.d. with mean μ and variance σ^2 . Then

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \rightsquigarrow \mathcal{N}(0, 1),$$

where \rightsquigarrow denotes convergence in distribution.

Convergence in distribution means that for all a, b we have

$$\mathbb{P}\left(a \leq \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \leq b\right) \rightarrow \mathbb{P}(a \leq Z \leq b) \quad \text{as } n \rightarrow \infty,$$

where $Z \sim \mathcal{N}(0, 1)$.

Remark.

Note that $(\sqrt{n}/\sigma)(\bar{X}_n - \mu)$ itself need not be normally distributed. For example if X_i is Bernoulli, then it takes value 0 or 1, so \bar{X}_n will take values in a discrete range: $\{0, 1/n, 2/n, \dots, (n-1)/n, 1\}$, whereas the normal distribution has continuous range.

Remark.

If $(\sqrt{n}/\sigma)(\bar{X}_n - \mu) \approx \mathcal{N}(0, 1)$ then

$$\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right). \tag{3}$$

This is the form in which we'll typically use the CLT. It tells us that

$$\mu - 3\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + 3\frac{\sigma}{\sqrt{n}}$$

with probability approximately 0.997. But we can also flip the inequality around to get

$$\bar{X}_n - 3\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 3\frac{\sigma}{\sqrt{n}},$$

also with probability 0.997! We have constructed our first confidence interval: in other words, we have given a probable range of values for μ .

We'll cover confidence intervals in more detail later on.

The Kiss Example

Do people prefer to turn their heads to the right when they kiss?

In an experiment in *Nature* [?], $n = 124$ couples were observed kissing. 80 of the couples turned their heads to the right when they kissed. That's a proportion of $80/124 = 0.645$, which is bigger than 0.5. Can we conclude for sure that humans have a preference to turn to the right? In other words — is 0.645 really “much bigger” than 0.5? Statistics will help us make this quantitative.

We model the n couples as $X_i \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(p)$, $i = 1, \dots, n$, where “Ber” stands for Bernoulli. Specifically, we let

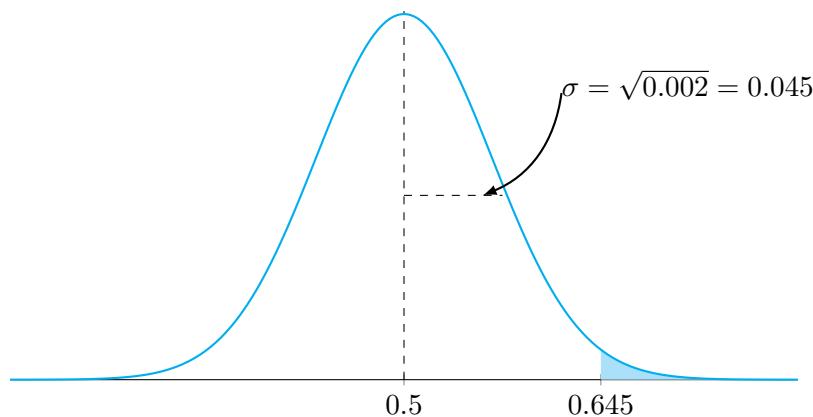
$$X_i = \begin{cases} 1 & \text{if turned right,} \\ 0 & \text{if turned left.} \end{cases}$$

Note that $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ is precisely the proportion of couples who turned their heads to the right. We have observed $\bar{X}_n = 0.645$.

We now want to know: what is the probability of observing $\bar{X}_n = 0.645$ given that $p = 1/2$? If this probability is sizeable, then it is reasonable that $p = 1/2$ is the true value of p which generated the data $X_i \sim \text{Ber}(p)$, and we can't conclude that there is a tendency to kiss turning your head to the right. But if the probability is very small, then we can confidently conclude that a right-turning preference does exist.

So let's do this computation: first we need the mean and variance of the X_i . The mean of $\text{Ber}(p)$ is p and the variance is $p(1-p)$, so if $p = 1/2$ then $\mu = \mathbb{E}[X_i] = 1/2$ and $\sigma^2 = \mathbb{V}[X_i] = 1/4$. By the CLT, we then have

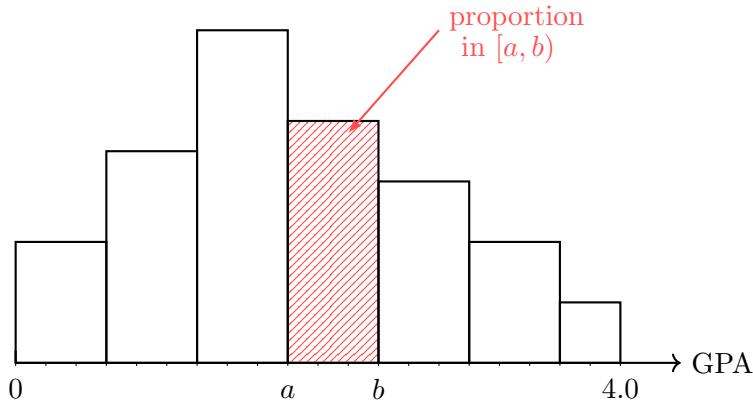
$$\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) = \mathcal{N}\left(1/2, \frac{1/4}{124}\right) = \mathcal{N}(0.5, 0.002).$$



We get $\mathbb{P}(\bar{X}_n \geq 0.645) \approx \mathbb{P}(\mathcal{N}(0.5, 0.002) \geq 0.645) \approx 0.003$. This is what's known as a p-value. Since it's tiny, we can be very confident that $1/2$ is *not* the right value, and that the true value of p is bigger than $1/2$ (meaning, there *is* a preference to turn your head to the right).

4 Basic data visualization: the histogram

Suppose x_1, \dots, x_n are the GPAs of the students in this class (we have $n = 233$). We can visualize the distribution of GPAs with a histogram.



- The area of the rectangle above $[a, b)$ is the proportion of GPAs between a and b :

$$\text{area} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(a \leq x_i < b), \quad (4)$$

where $\mathbb{1}(\cdot)$ is called an *indicator* function. It evaluates to 1 if the statement inside the parentheses is true and it evaluates to 0 if the statement is false.

- Since $\text{area} = (b - a) \times \text{height}$, we get the height of the column by dividing the area by $b - a$.
- Caution:** *only if* the bins are equally spaced can we visually judge the proportions of GPAs in each bin by looking at the heights. If the bins are *not* equally spaced then the heights don't tell us everything (a column could be unusually tall if it corresponds to a very small bin size.)

4.1 Shapes

We can also smooth out the histogram with a “kernel density estimator” (KDE), which we'll learn about in May. A smoothed out histogram tells us about the *shape* of the distribution; see Figure 1.

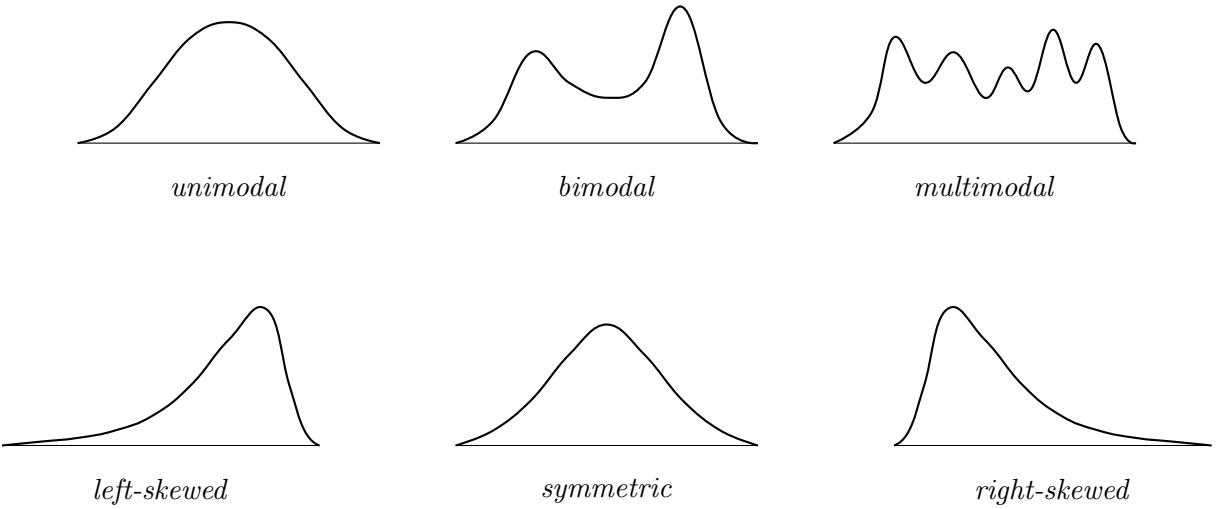


Figure 1: Shapes a distribution can take. Is there skew? Are there one or several modes?

5 Summary statistics

We can summarize the data with a few summary statistics:

- **mean**
- **standard deviation**
- **median**
 - splits the data in half:
half of the data points x_1, \dots, x_n are to the left and half to the right
 - Formally: $\frac{1}{n} \sum_{i=1}^n \mathbb{1}(x_i \leq \text{median}) = 1/2$.
- **quantiles**
 - a generalization of the median
 - **the quantile of order $1 - \alpha$, or $100(1 - \alpha)$ percentile, denoted q_α**
is a number such that α of the data is *above* it.
 - Formally: $\frac{1}{n} \sum_{i=1}^n \mathbb{1}(x_i \leq q_\alpha) = 1 - \alpha$.
 - Important special cases:
1st quartile $Q_1 = q_{0.75}$ (3/4 of the data is to the right),
3rd quartile $Q_3 = q_{0.25}$ (1/4 of the data is to the right).
 - Note that $Q_1 < \text{median} < Q_3$. The median is the second quartile.
- **interquartile range IQR** = $Q_3 - Q_1$.

Figure 2 shows some of these statistics on a smoothed histogram.

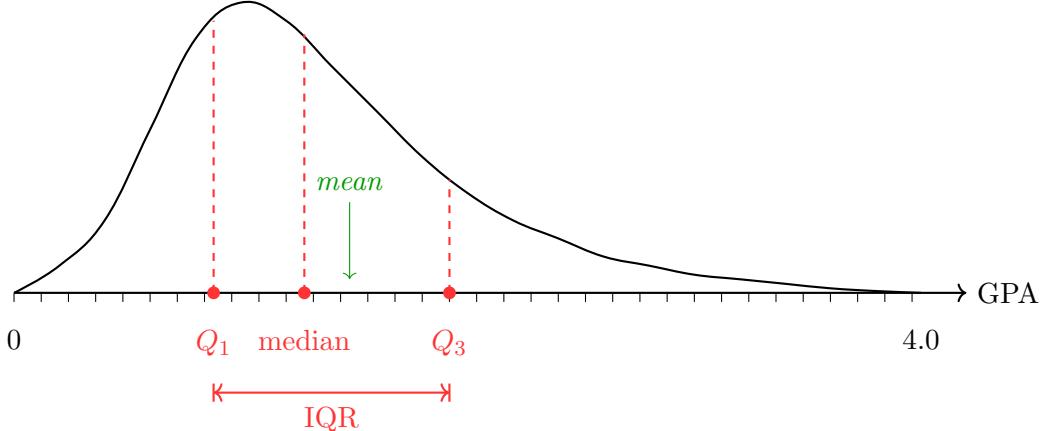


Figure 2: A smoothed histogram showing locations of the summary statistics. Note that the mean is to the right of the median because the distribution is right-skewed.

5.1 Robustness

Outliers are abnormally large or small values compared to the rest of the data. Formally:

$$x_i \text{ is an outlier if } x_i > Q_3 + 1.5\text{IQR} \text{ or } x_i < Q_1 - 1.5\text{IQR}.$$

The mean and standard deviation are strongly affected by outliers. For example, very large outliers pull the mean to the right of the median, as in Figure 2. On the other hand, the median and IQR are robust to outliers (if the largest data point is doubled, say, this will not change the location of the median and IQR). The following table summarizes four important summary statistics.

location	spread
mean	std. dev.
median	IQR \leftarrow robust

6 Visualizing summary statistics

Another way to concisely depict summary statistics is with a *box plot*, also sometimes called a “box and whiskers” plot; see Figure 3. The left and right endpoints of the box are Q_1 and Q_3 , respectively, and a line is drawn in between to denote the location of the median. The box is our visual representation for the middle 50% of the data, and the location of the median between Q_1 and Q_3 conveys whether the distribution is skewed in one direction.

The two line segments extending to the left and right from the box are the “whiskers”. The length of the whiskers is 1.5IQR, so that by definition, the outliers

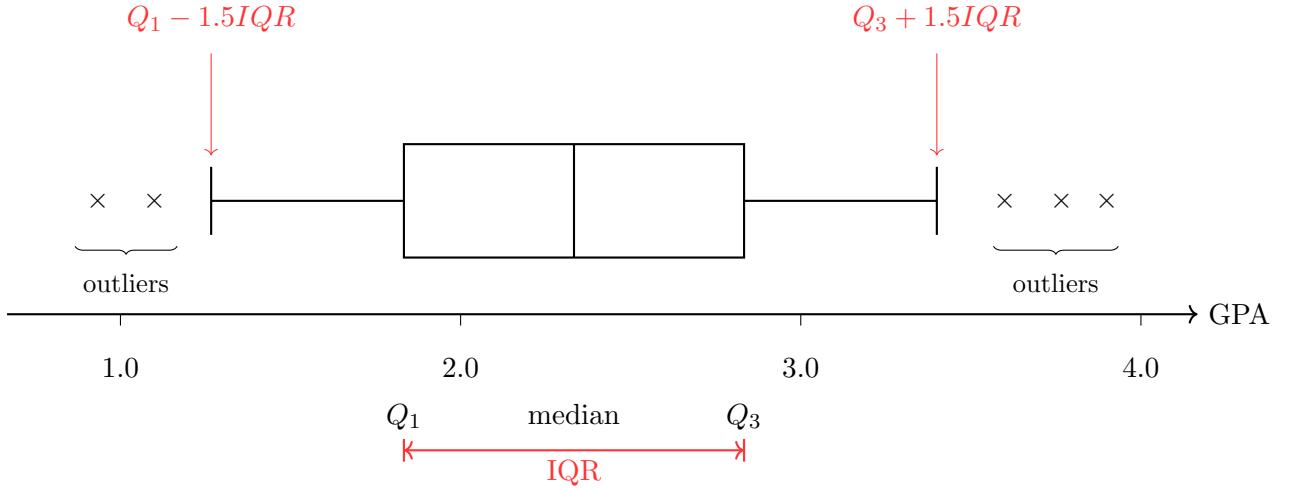


Figure 3: A box plot

are to the left of the left whisker and to the right of the right whisker. The locations of the outliers are indicated explicitly on the boxplot.

6.1 Data with multiple variables: scatterplot and comparative boxplot

So far we've only considered one-dimensional data. But we could also have, e.g., pairs (X_i, Y_i) . For example, X_i denotes the number of days a month a student smokes marijuana, and Y_i denotes the student's GPA. A common way to depict such data is with a scatterplot, as in Figure 4. We simply plot the location of each data point in the X - Y plane.

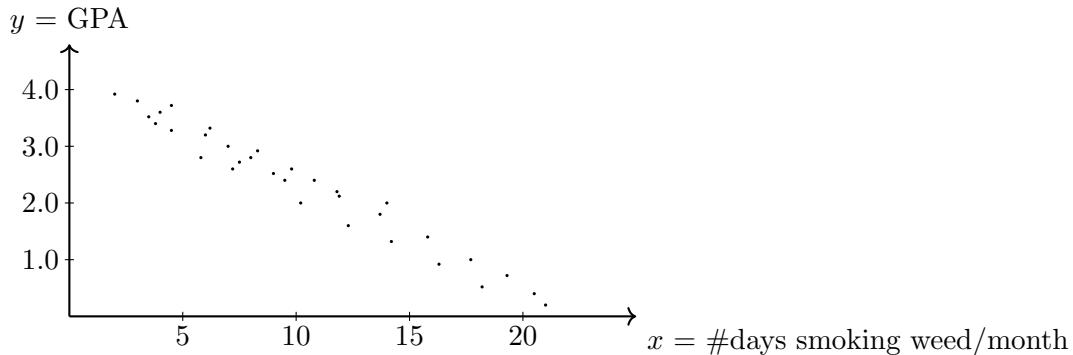


Figure 4: A scatterplot

If the X -values tend to be clustered or if the X -values are not numbers at all (e.g. $X_i \in \{\text{freshman, sophomore, junior, senior}\}$), then we can depict the data using several boxplots for the Y distributions, one for each cluster/category of X

values. This lets us visualize the difference in the Y distributions across different X values. This is called a comparative boxplot; see Figure 5.

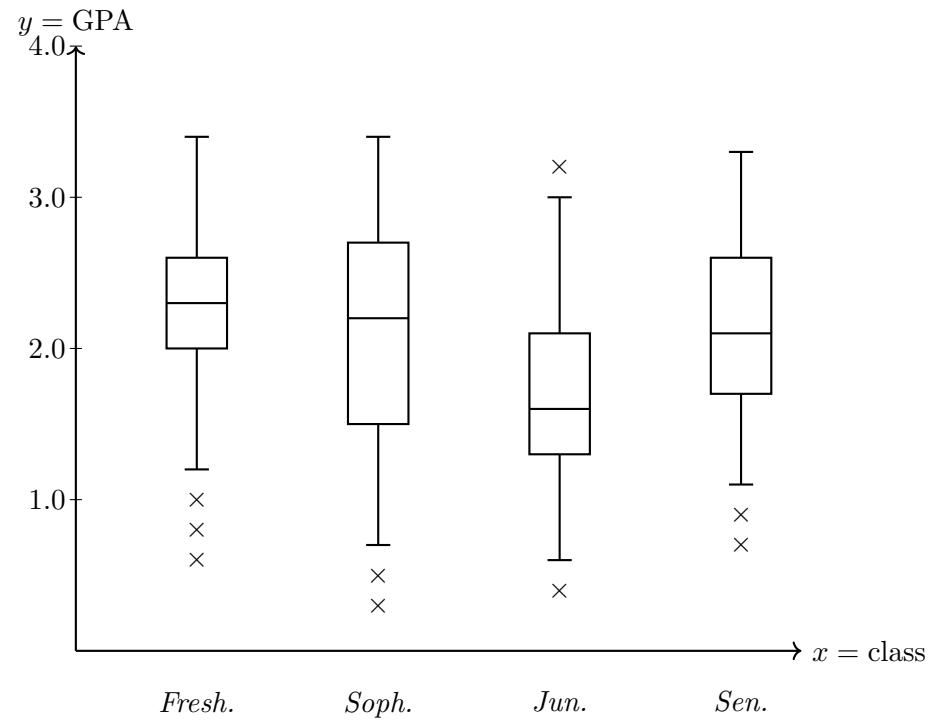


Figure 5: Comparative boxplots