

Statistics 1 Notes

paraphrased by Tyler Wright

*An important note, these notes are absolutely **NOT** guaranteed to be correct, representative of the course, or rigorous. Any result of this is not the author's fault.*

1 The Basics of Data Analysis

1.1 Samples

A sample is a set of values observed from a simple random sample of some size n from a population where each sample member is chosen **independently** of each other and each population member is **equally likely** to be selected.

Samples are usually written as $\{x_1, x_2, \dots, x_n\}$ where each x_i represents an observed value. If the data is ordered, the data is written as $\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$ (for numerical values, this is ascending order). So, in this case, $x_{(1)}$ is always the minimum, $x_{(n)}$ is always the maximum.

1.2 Probability Density Functions

For a sample $\{x_1, x_2, \dots, x_n\}$, we can imagine each datum as being distributed with some population distribution X . As each datum is independent of all other observed values, we can write the probability density of this sample as follows:

$$f_X(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_X(x_i).$$

1.3 Measures of Central Tendency

1.3.1 Sample Median

For a sample $X = \{x_1, x_2, \dots, x_n\}$, we define the sample median M as follows:

$$M(X) = \begin{cases} x_{(m+1)} & \text{for } n = 2m + 1 \\ \frac{x_{(m)} + x_{(m+1)}}{2} & \text{for } n = 2m. \end{cases}$$

Essentially, it equals the middle value or the average of the middle values. Also, it's important to note that the median is not sensitive to extreme values.

1.3.2 Sample Mean

For a sample $X = \{x_1, x_2, \dots, x_n\}$, we define the sample mean \bar{X} as follows:

$$\bar{X} = \frac{1}{n} \left(\sum_{i=1}^n x_i \right).$$

This is easy to calculate even when combining samples. However, it is sensitive to extreme values.

1.3.3 Trimmed Sample Mean

For a sample $X = \{x_1, x_2, \dots, x_n\}$, we define the trimmed sample mean \overline{X}_Δ for some percentage $\Delta\%$ as follows:

$$\begin{aligned}\text{Let } k &= \left\lfloor n \frac{\Delta}{100} \right\rfloor \\ \text{Let } \tilde{X} &= \{x_{(k+1)}, x_{(k+2)}, \dots, x_{(n-k)}\} \\ \overline{X}_\Delta &= \overline{\tilde{X}} \text{ (the sample mean of } \tilde{X}\text{).}\end{aligned}$$

Basically, you remove the first and last $\Delta\%$ of values and take the sample mean of the remaining values.

1.4 Measures of Spread

1.4.1 Sample Variance

For a sample $X = \{x_1, x_2, \dots, x_n\}$, we define the sample variance s^2 as follows:

$$\begin{aligned}s^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right).\end{aligned}$$

This measures how much the data varies.

1.4.2 Hinges

There are two hinge measures, lower (H_1) and upper (H_3):

$$\begin{aligned}H_1 &= \text{median of \{data values } \leq \text{ the median\}} \\ H_3 &= \text{median of \{data values } \geq \text{ the median\}}.\end{aligned}$$

1.4.3 Quartiles

For a sample $X = \{x_1, x_2, \dots, x_n\}$, there are two quartile measures, lower (Q_1) and upper (Q_3). The formulas are long and overly complicated so for Q_1 :

- Calculate $k = \frac{n+1}{4}$
- If $k \in \mathbb{Z}$, $Q_1 = x_{(k)}$
- Otherwise, do linear interpolation between $x_{(\lfloor k \rfloor)}$ and $x_{(\lfloor k \rfloor + 1)}$

And similarly for Q_3 :

- Calculate $k = 3 \left(\frac{n+1}{4} \right)$
- If $k \in \mathbb{Z}$, $Q_3 = x_{(k)}$
- Otherwise, do linear interpolation between $x_{(\lfloor k \rfloor)}$ and $x_{(\lfloor k \rfloor + 1)}$

For large samples, the quartiles and hinges tend to be close to each other.

1.4.4 Interquartile Range (IQR)

The IQR is the difference between Q_3 and Q_1 ($Q_3 - Q_1$).

In this course, outliers are defined as more than $\frac{3}{2}(\text{IQR})$ (or approx. $\frac{3}{2}(H_3 - H_1)$) from the median.

1.4.5 Skewness

We measure skewness by the distance of the hinges from the median. If H_3 is further from the median than H_1 , we have a longer right tail. If the converse is true, we have a longer left tail.

2 Assessing Fit

2.1 Quantiles of a Distribution

For a distribution X with cumulative distribution function F_X , the quantiles of the distribution are defined as the set of values:

$$F_X^{-1} \left\{ \frac{1}{n+1}, \dots, \frac{n}{n+1} \right\}.$$

We use $n+1$ on the denominator as $F_X^{-1}(1)$ can be ∞ .

The ordered sample is called the set of sample quantiles.

2.2 Quantile-Quantile (Q-Q) Plots

These are the steps for constructing a Q-Q plot of a sample $\{x_1, x_2, \dots, x_n\}$ with cumulative distribution function F_X :

- Generate an estimate for the parameter(s) $(\hat{\theta}_1, \hat{\theta}_2, \dots)$
- Compute the quantiles (the expected quantiles if the hypothesised model is correct)
- Plot each expected quantile against the sample quantile $(F_X^{-1}(\frac{k}{n+1}; \hat{\theta}), x_{(k)})$.

What we would expect, if our hypothesis is correct, is that the plotted points lie close to the line $y = x$. This is saying our sample and expected quantiles are close together.

2.3 Probability Plots

These are similar to the Q-Q plots but plot the sample cumulative probability against expected probability $(F_X(x_{(k)}), \frac{k}{n+1})$.

3 Estimation

We have that a population is distributed with some distribution X with a probability density function (PDF) f_X , cumulative distribution function (CDF) F_X , and some parameters $\{\theta_1, \dots\}$. We can make guesses at the distribution of a sample and use tests to verify that. But, to do these tests we need a value for the parameters. It's not practical to guess these, so we need to estimate them.

3.1 Parameters

We say $\hat{\theta}$ is an estimator for θ and define it as a function of a sample $\{x_1, x_2, \dots, x_n\}$:

$$\hat{\theta}(x_1, x_2, \dots, x_n).$$

3.2 Distribution Quantities

From our estimated value of the distribution parameters, we can calculate estimated values for distribution quantities like the mean and variance. We consider τ a function of the parameter that gives a distribution quantity:

- **True quantity:** $\tau(\theta)$ where θ is the true distribution parameter
- **Estimated quantity:** $\hat{\tau} = \tau(\hat{\theta})$ where $\hat{\theta}$ is our estimated parameter.

4 Method of Moments Estimation

4.1 Definition of a Moment

The k th moment of a probability distribution X is defined as follows:

$$\mathbb{E}(X^k) := \int_{-\infty}^{\infty} x^k f_X(x) dx.$$

Setting $k = 1$ gives us the familiar expectation of X :

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

In the discrete case, the integral is a sum.

We define the k th sample moment m_k as follows:

$$m_k = \frac{\sum_{i=1}^n x_i^k}{n}.$$

Or rather, the k th moment is the average value of x^k in the sample.

4.2 The Process

By considering the a probability distribution with parameter θ , we can find functions for the moments in terms of θ . These can be rearranged to give functions for θ in terms of the moments. We can then use the sample moments to generate an estimate for θ .

4.3 Method of Moments on the Exponential

Assume we have some population X distributed according to the Exponential with some parameter θ . We say $X \sim \text{Exp}(\theta)$.

$$\begin{aligned} f_X(x) &= \theta e^{-\theta x} && (x > 0) \\ \Rightarrow \mathbb{E}(X) &= \frac{1}{\theta} \\ \Rightarrow \theta &= \frac{1}{\mathbb{E}(X)} \\ \Rightarrow \hat{\theta}_{mom} &= \frac{1}{m_1}. \end{aligned}$$

If there were more parameters, we would have to consider greater moments of X .

5 Maximum Likelihood Estimation