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 $\{x_1, x_2, \ldots, x_n\}$ where each x_i represents an observed value. If the data is ordered, the data is written as $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ (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 $\{x_1, x_2, \ldots, x_n\}$, 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:

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

1.3 Measures of Central Tendency

1.3.1 Sample Median

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

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

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

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

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 $X = \{x_1, x_2, \ldots, x_n\}$, 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):

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

1.4.3 Quartiles

For a sample $X = \{x_1, x_2, \ldots, x_n\}$, 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 $x_{(\lfloor k+1 \rfloor)}$

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 $x_{(\lfloor k+1 \rfloor)}$

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 $\frac{3}{2}$(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:

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

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 $\{x_1, x_2, \ldots, x_n\}$ with cumulative distribution function F_X:

- Generate an estimate for the parameter(s) $(\hat{\theta}_1, \hat{\theta}_2, \ldots)$
- 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 $\{\theta_1, \ldots\}$. We can make guesses at the distribution of a sample and use tests to verify that. But, to do these tests we need a valu 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 $\{x_1, x_2, \ldots, x_n\}$:

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

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 kth 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:

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

In the discrete case, the integral is a sum.

We define the kth sample moment m_k as follows:

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

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

4.2 The Process

By considering the a probability distribution X with parameter θ, we can find functions for the moments of X 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 ($\hat{\theta}_{mom}$).

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 $X \sim \text{Exp}(\theta)$.

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

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

5 Maximum Likelihood Estimation

5.1 The Process

By considering the a probability distribution X with parameter θ, we can find functions for the probability of events occuring in terms of θ. If we find where this function is maximised, it will give us the value of θ that makes this sample most likely. This is the maximum likelihood estimate ($\hat{\theta}_{mle}$).

5.2 Optimisation of the Method

Consider a sample $\{x_1, x_2, \ldots, x_n\}$ with distributions $\{X_1, X_2, \ldots, X_n\}$, we call the likelihood function the joint PDF of $\{X_1, X_2, \ldots, X_n\}$. We input our sample values ($L = f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$) which gives us a function in terms of our unknown parameters $\theta_1, \theta_2, \ldots$.

For $X_1, X_2, \ldots, X_n$ independent and identically distributed sharing some distribution X, the joint PDF can be written as a product of marginals:

$$\begin{aligned} L &= f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) \\ &= f_X(x_1) f_X(x_2) \cdots f_X(x_n) \\ &= \prod_{i=1}^{n} f_X(x_i). \end{aligned}$$

We can take the natural logarithm of this likelihood function (the value where it's maximised is preserved as the natural logarithm is increasing) ($\ell = ln(L)$). If, again, $X_1, X_2, \ldots, X_n$ are independent and identically distributed sharing some distribution X:

$$\begin{aligned} \ell &= \ln\left(\prod_{i=1}^{n} f_X(x_i)\right) \\ &= \sum_{i=1}^{n} \left[\ln\left(f_X(x_i)\right)\right] \end{aligned}$$

We know $\hat{\theta}_{mle}$ is the solution to:

$$\frac{\partial}{\partial\theta}\ell(\theta) = 0.$$

So, in the independent and identically distributed case:

$$\frac{\partial}{\partial\theta}\ell(\theta) = 0 \iff \sum_{i=1}^{n}\left[\frac{\partial}{\partial\theta}\ln\left(f_X(x_i)\right)\right] = 0.$$

5.3 Multiple Parameters

When finding the maximum likelihood estimate for multiple parameters, we obtain multiple equations by partially differentiating ℓ by our different parameters, giving an equation for each.

5.4 Non-regular density

If our function L is piecewise, we may find that our maximum isn't where the derivative is zero, but as the endpoints of the parts of the function.

Consider the maximum of $f : \mathbb{R} \to \mathbb{R}$ where:

$$f(x) = \begin{cases} \theta^{-x} & \text{for } x \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

6 The Performance of Estimators

6.1 Variation of Estimators

We can consider the distribution of an estimator to compare them. We consider these main quantities:

- $\text{bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$
- $\text{mse}(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \theta)^2)$.

Bias is the quantity we expect our estimator to vary by from the true value. The MSE (mean squared error) is how much is varies.

We can rewrite the formula for the mean squared error as follows:

$$\text{mse}(\hat{\theta}) = \text{Var}(\hat{\theta}) - \text{bias}(\hat{\theta}).$$

6.2 Method of Simulation

If we have a distribution with known parameters $\theta_1, \theta_2, \ldots$, we can sample N samples of size n and use our estimators to calculate estimates for these known parameters for each sample.

From this, we can calculate the average error, sample variance, and average squared error of each estimator. These quantities estimate bias, variance, and mean squared error respectively.

If we repeat this process for multiple estimators, we can compare our estimators with these quantities.

7 Central Limit Theorem

7.1 Definition of the Central Limit Theorem

For $X_1, X_2, \ldots, X_n$ a random sample from a population with mean $\mu = \mathbb{E}(X)$ and variance $\sigma^2 = \mathrm{Var}(X)$. Let $\overline{X_n}$ be the sample mean. For n large we have:

$$\mathbb{P}\left(\sqrt{n}\left[\frac{\overline{X_n} - \mu}{\sigma}\right] \leq x\right) \simeq \mathbb{P}(\mathcal{N}(0,1) \leq x) = \Phi(x).$$

Or similarly:

$$\overline{X_n} \simeq \mathcal{N}(\mu, \sigma^2/n).$$

7.2 Continuity Correction

When using the Central Limit Theorem to approximate discrete random variables, it is important to make a continuity correction. Let $X_1, X_2, \ldots, X_n$ be samples from a discrete random variable with sample mean $\overline{X_n}$, population mean μ, and population variance σ^2:

$$\mathbb{P}\left(\sqrt{n}\left[\frac{\overline{X_n} - \mu}{\sigma}\right] \leq x\right) \simeq \mathbb{P}(\mathcal{N}(0,1) < x + \frac{1}{2})$$
$$\mathbb{P}\left(\sqrt{n}\left[\frac{\overline{X_n} - \mu}{\sigma}\right] < x\right) \simeq \mathbb{P}(\mathcal{N}(0,1) < x - \frac{1}{2})$$

8 A Reminder on Moment Generating Functions

8.1 Definition of a Moment Generating Function (MGF)

For a random variable X, we define the moment generating function by:

$$\mathcal{M}_X(t) := \mathbb{E}(e^{tX}) = \begin{cases} \int_{-\infty}^{\infty} e^{tX} f_X(x)\,dx & \text{for } X \text{ continuous} \\ \sum_{x \in S} e^{tX} \mathbb{P}(X = x) & \text{for } X \text{ discrete.} \end{cases}$$

8.2 Properties of a Moment Generating Function

8.2.1 Standard examples of moment generating functions

For a random variable X:

- $X \sim \mathcal{N}(\mu, \sigma^2) \Leftrightarrow \mathcal{M}_X(t) = \exp\left(\mu t + \frac{(\sigma t)^2}{2}\right)$
- $X \sim \text{Exp}(\theta) \Leftrightarrow \mathcal{M}_X(t) = \frac{\theta}{\theta - t}$
- $X \sim \text{Gamma}(\alpha, \beta) \Leftrightarrow \mathcal{M}_X(t) = \frac{\beta^\alpha}{(\beta - t)^\alpha}$.

8.2.2 Joint moment generating functions

The joint MGF of X and Y is:

$$\mathcal{M}_{X,Y}(s,t) := \mathbb{E}(e^{sX+tY}).$$

They are such that:

$$\begin{aligned} \mathcal{M}_X(s) &= \mathcal{M}_{X,Y}(s,0) \\ \mathcal{M}_Y(t) &= \mathcal{M}_{X,Y}(0,t). \end{aligned}$$

We also have that X and Y are independent if and only if:

$$\mathcal{M}_{X,Y}(s,t) = \mathcal{M}_X(s)\mathcal{M}_Y(t).$$

8.2.3 Independence of moment generating functions

If $X_1, X_2, \dots, X_n$ are independent and $Y = \sum_{i=1}^n X_i$:

$$\mathcal{M}_Y(t) = \prod_{i=1}^n \mathcal{M}_{X_i}(t)$$

8.2.4 Uniqueness of moment generating functions

The MGF uniquely defines a distribution, for two random variables X, Y:

$$\mathcal{M}_X = \mathcal{M}_Y \Leftrightarrow X = Y.$$

9 The Normal Distribution

9.1 Transformation and Addition of the Normal

For $X \sim \mathcal{N}(\mu, \sigma^2)$ and $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i \in \{1, 2, \ldots, n\}$, let $\overline{X} = \frac{1}{n}(\sum_{i=1}^n X_i)$ be the sample mean:

$$aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2) \tag{Linear Transformation}$$

$$\sum_{i=1}^n X_i \sim \mathcal{N}(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2) \tag{Summed}$$

$$\frac{(X - \mu)}{\sigma} \sim \mathcal{N}(0, 1) \tag{Standardised}$$

$$\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \tag{Sample Mean}$$

$$\sqrt{n}\left(\frac{\overline{X} - \mu}{\sigma}\right) \sim \mathcal{N}(0, 1). \tag{Standardised Sample Mean}$$

It's very important to remember that multiplication and summing differ when dealing with the Normal (when it comes to the variance). So, if you have a Normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, $2X \neq X + X$ as $\mathrm{Var}(2X) = 4\sigma^2$ and $\mathrm{Var}(X + X) = 2\sigma^2$. This is because when you're multiplying, you're amplifying variation in your sample, but when you sum you're combining variance.

9.2 Independence of the Sample Mean and the Sum of Squared Difference

For $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i \in \{1, 2, \ldots, n\}$, let $\overline{X} = \frac{1}{n}(\sum_{i=1}^n X_i)$ be the sample mean. We have that $\overline{X}$ and $\sum_{i=1}^n (X_i - \overline{X})^2$ are independent.

10 Sampling Distributions related to the Normal

10.1 The χ^2 Distribution

10.1.1 Definition of the χ^2 distribution

We say that a random variable $X \sim \chi_r^2$ (r degrees of freedom) if:

$$\mathcal{M}_X(t) = (1 - 2t)^{-r/2}.$$

10.1.2 Properties of the χ^2 distribution

Let $X \sim \chi^2_r$, $Y \sim \chi^2_s$:

- $X \sim \Gamma(\frac{r}{2}, \frac{1}{2})$
- $\mathbb{E}(X) = r$
- $\text{Var}(X) = 2r$
- $X + Y \sim \chi^2_{r+s}$

We also have some results relating to the Normal, let Z be the standard Normal, X_i for $i \in \{1, 2, \ldots, n\}$ be samples from $\mathcal{N}(\mu, \sigma^2)$, let $\overline{X} = \frac{1}{n}(\sum_{i=1}^n X_i)$ be the sample mean:

- $Z^2 \sim \chi^2_1$
- $\sum_{i=1}^n (\frac{X_i - \mu}{\sigma})^2 \sim \chi^2_n$
- $\sum_{i=1}^n (\frac{X_i - \overline{X}}{\sigma})^2 \sim \chi^2_{n-1}$.

Finally, we also have some results relating to the Exponential and Gamma distributions, let X_i for $i \in \{1, 2, \ldots, n\}$ be samples from $\text{Exp}(\theta)$:

- $\sum_{i=1}^n X_i \sim \Gamma(n, \theta)$
- $\overline{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim \Gamma(n, n\theta)$
- $2\theta \sum_{i=1}^n X_i \sim \Gamma(n, 1/2) = \chi^2_{2n}$

10.2 The t Distribution

10.2.1 Definition of the t distribution

For $Z \sim \mathcal{N}(0, 1)$, $X \sim \chi^2_r$ **independent** we have:

$$T = \frac{Z}{\sqrt{X/r}},$$

is distributed with a t distribution with r degrees of freedom (t_r).

10.2.2 Properties of the t distribution

For $T \sim t_r$:

- $\mathbb{E}(T) = 0$
- $\text{Var}(T) = \frac{r}{r-2}$
- The density of T approaches $\mathcal{N}(0, 1)$ as $r \to \infty$.

10.2.3 Samples from the Normal with σ unknown

For X_i for $i \in \{1, 2, \ldots, n\}$ be samples from $\mathcal{N}(\mu, \sigma^2)$, let $\overline{X} = \frac{1}{n}(\sum_{i=1}^{n} X_i)$ be the sample mean and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ be the sample variance. We have that:

$$\sqrt{n}\left(\frac{\overline{X} - \mu}{S}\right) \sim t_{n-1}$$

*This is **extremely** key as this allows us to perform hypothesis test on any Normal sample without knowing the population variance.*