VIP Cheatsheet: Statistics

Afshine Amidi and Shervine Amidi

September 8, 2020

Paramater estimation

❒ **Random sample** – A random sample is a collection of n random variables $X_1, ..., X_n$ that are independent and identically distributed with X.

❒ **Estimator** – An estimator $\hat{\theta}$ is a function of the data that is used to infer the value of an unknown parameter θ in a statistical model.

❒ **Bias** – The bias of an estimator $\hat{\theta}$ is defined as being the difference between the expected value of the distribution of $\hat{\theta}$ and the true value, i.e.:

$$\boxed{\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta}$$

Remark: an estimator is said to be unbiased when we have $E[\hat{\theta}] = \theta$.

❒ **Sample mean and variance** – The sample mean and the sample variance of a random sample are used to estimate the true mean μ and the true variance σ^2 of a distribution, are noted $\overline{X}$ and s^2 respectively, and are such that:

$$\boxed{\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i} \quad \text{and} \quad \boxed{s^2 = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

❒ **Central Limit Theorem** – Let us have a random sample $X_1, ..., X_n$ following a given distribution with mean μ and variance σ^2, then we have:

$$\boxed{\overline{X} \underset{n\to+\infty}{\sim} \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)}$$

Confidence intervals

❒ **Confidence level** – A confidence interval $CI_{1-\alpha}$ with confidence level $1-\alpha$ of a true parameter θ is such that $1-\alpha$ of the time, the true value is contained in the confidence interval:

$$\boxed{P(\theta \in CI_{1-\alpha}) = 1 - \alpha}$$

❒ **Confidence interval for the mean** – When determining a confidence interval for the mean μ, different test statistics have to be computed depending on which case we are in. The following table sums it up:

Distribution	Sample size	σ^2	Statistic	$1-\alpha$ confidence interval
$X_i \sim \mathcal{N}(\mu, \sigma)$	any	known	$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$	$\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$
	small	unknown	$\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$	$\left[\overline{X} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \overline{X} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$
$X_i \sim$ any	large	known	$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$	$\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$
		unknown	$\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} \sim \mathcal{N}(0,1)$	$\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$
$X_i \sim$ any	small	any	Go home!	Go home!

❒ **Confidence interval for the variance** – The single-line table below sums up the test statistic to compute when determining the confidence interval for the variance.

Distribution	Sample size	μ	Statistic	$1-\alpha$ confidence interval
$X_i \sim \mathcal{N}(\mu,\sigma)$	any	any	$\frac{s^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$	$\left[\frac{s^2(n-1)}{\chi^2_2}, \frac{s^2(n-1)}{\chi^2_1}\right]$

Hypothesis testing

❒ **Errors** – In a hypothesis test, we note α and β the type I and type II errors respectively. By noting T the test statistic and R the rejection region, we have:

$$\boxed{\alpha = P(T \in R | H_0 \text{ true})} \quad \text{and} \quad \boxed{\beta = P(T \notin R | H_1 \text{ true})}$$

❒ **p-value** – In a hypothesis test, the p-value is the probability under the null hypothesis of having a test statistic T at least as extreme as the one that we observed T_0. We have:

Case	Left-sided	Right-sided	Two-sided
p-value	$P(T \leqslant T_0 \| H_0 \text{ true})$	$P(T \geqslant T_0 \| H_0 \text{ true})$	$P(\|T\| \geqslant \|T_0\| \| H_0 \text{ true})$

❒ **Sign test** – The sign test is a non-parametric test used to determine whether the median of a sample is equal to the hypothesized median. By noting V the number of samples falling to the right of the hypothesized median, we have:

Statistic when $np < 5$	Statistic when $np \geqslant 5$
$V \underset{H_0}{\sim} \mathcal{B}\left(n, p = \frac{1}{2}\right)$	$Z = \frac{V - \frac{n}{2}}{\frac{\sqrt{n}}{2}} \underset{H_0}{\sim} \mathcal{N}(0,1)$

❒ **Testing for the difference in two means** – The table below sums up the test statistic to compute when performing a hypothesis test where the null hypothesis is:

$$H_0 \quad : \quad \mu_X - \mu_Y = \delta$$

Distribution of X_i, Y_i	n_X, n_Y	σ_X^2, σ_Y^2	Statistic
Normal	any	known	$\frac{(\overline{X}-\overline{Y})-\delta}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}} \underset{H_0}{\sim} \mathcal{N}(0,1)$
	large	unknown	$\frac{(\overline{X}-\overline{Y})-\delta}{\sqrt{\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}}} \underset{H_0}{\sim} \mathcal{N}(0,1)$
	small	unknown $\sigma_X = \sigma_Y$	$\frac{(\overline{X}-\overline{Y})-\delta}{s\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}} \underset{H_0}{\sim} t_{n_X+n_Y-2}$
Normal, paired $D_i = X_i - Y_i$	any $n_X = n_Y$	unknown	$\frac{\overline{D}-\delta}{\frac{s_D}{\sqrt{n}}} \underset{H_0}{\sim} t_{n-1}$

❒ **χ^2 goodness of fit test** – By noting k the number of bins, n the total number of samples, p_i the probability of success in each bin and Y_i the associated number of samples, we can use the test statistic T defined below to test whether or not there is a good fit. If $np_i \geqslant 5$, we have:

$$T = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i} \underset{H_0}{\sim} \chi^2_{df} \quad \text{with} \quad df = (k-1) - \#(\text{estimated parameters})$$

❒ **Test for arbitrary trends** – Given a sequence, the test for arbitrary trends is a non-parametric test, whose aim is to determine whether the data suggest the presence of an increasing trend:

$$H_0 : \text{no trend} \quad \text{versus} \quad H_1 : \text{there is an increasing trend}$$

If we note x the number of transpositions in the sequence, the p-value is computed as:

$$p\text{-value} = P(T \leqslant x)$$

Regression analysis

In the following section, we will note $(x_1, Y_1), ..., (x_n, Y_n)$ a collection of n data points.

❒ **Simple linear model** – Let X be a deterministic variable and Y a dependent random variable. In the context of a simple linear model, we assume that Y is linked to X via the regression coefficients α, β and a random variable $e \sim \mathcal{N}(0,\sigma)$, where e is referred as the error. We estimate Y, α, β by $\hat{Y}, A, B$ and have:

$$Y = \alpha + \beta X + e \quad \text{and} \quad \hat{Y}_i = A + Bx_i$$

❒ **Notations** – Given n data points (x_i, Y_i), we define S_{XY},S_{XX} and S_{YY} as follows:

$$S_{XY} = \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) \quad \text{and} \quad S_{XX} = \sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \text{and} \quad S_{YY} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

❒ **Sum of squared errors** – By keeping the same notations, we define the sum of squared errors, also known as SSE, as follows:

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (A + Bx_i))^2 = S_{YY} - BS_{XY}$$

❒ **Least-squares estimates** – When estimating the coefficients α, β with the least-squares method which is done by minimizing the SSE, we obtain the estimates A, B defined as follows:

$$A = \overline{Y} - \frac{S_{XY}}{S_{XX}}\overline{x} \quad \text{and} \quad B = \frac{S_{XY}}{S_{XX}}$$

❒ **Key results** – When σ is unknown, this parameter is estimated by the unbiased estimator s^2 defined as follows:

$$s^2 = \frac{S_{YY} - BS_{XY}}{n-2} \quad \text{and we have} \quad \frac{s^2(n-2)}{\sigma^2} \sim \chi^2_{n-2}$$

The table below sums up the properties surrounding the least-squares estimates A, B when σ is known or not:

Coeff	σ	Statistic	$1-\alpha$ confidence interval
α	known	$\frac{A-\alpha}{\sigma\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}}} \sim \mathcal{N}(0,1)$	$\left[A - z_{\frac{\alpha}{2}}\sigma\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}}, A + z_{\frac{\alpha}{2}}\sigma\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}}\right]$
	unknown	$\frac{A-\alpha}{s\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}}} \sim t_{n-2}$	$\left[A - t_{\frac{\alpha}{2}}s\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}}, A + t_{\frac{\alpha}{2}}s\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}}\right]$
β	known	$\frac{B-\beta}{\frac{\sigma}{\sqrt{S_{XX}}}} \sim \mathcal{N}(0,1)$	$\left[B - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{S_{XX}}}, B + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{S_{XX}}}\right]$
	unknown	$\frac{B-\beta}{\frac{s}{\sqrt{S_{XX}}}} \sim t_{n-2}$	$\left[B - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{S_{XX}}}, B + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{S_{XX}}}\right]$

Correlation analysis

❒ **Sample correlation coefficient** – The correlation coefficient is in practice estimated by the sample correlation coefficient, often noted r or $\hat{\rho}$, which is defined as:

$$r = \hat{\rho} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} \quad \text{with} \quad \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \underset{H_0}{\sim} t_{n-2} \quad \text{for } H_0 : \rho = 0$$

❒ **Correlation properties** – By noting $V_1 = V - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}}$, $V_2 = V + \frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}}$ with $V = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$, the table below sums up the key results surrounding the correlation coefficient estimate:

Sample size	Standardized statistic	$1-\alpha$ confidence interval for ρ
large	$\frac{V - \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)}{\frac{1}{\sqrt{n-3}}} \underset{n\gg 1}{\sim} \mathcal{N}(0,1)$	$\left[\frac{e^{2V_1}-1}{e^{2V_1}+1}, \frac{e^{2V_2}-1}{e^{2V_2}+1}\right]$