VIP Cheatsheet: Probability

Afshine Amidi and Shervine Amidi

September 8, 2020

Introduction to Probability and Combinatorics

❒ **Sample space** – The set of all possible outcomes of an experiment is known as the sample space of the experiment and is denoted by S.

❒ **Event** – Any subset E of the sample space is known as an event. That is, an event is a set consisting of possible outcomes of the experiment. If the outcome of the experiment is contained in E, then we say that E has occurred.

❒ **Axioms of probability** – For each event E, we denote $P(E)$ as the probability of event E occurring. By noting $E_1,...,E_n$ mutually exclusive events, we have the 3 following axioms:

$$(1)\quad \boxed{0 \leqslant P(E) \leqslant 1} \qquad (2)\quad \boxed{P(S)=1} \qquad (3)\quad \boxed{P\left(\bigcup_{i=1}^{n} E_i\right)=\sum_{i=1}^{n} P(E_i)}$$

❒ **Permutation** – A permutation is an arrangement of r objects from a pool of n objects, in a given order. The number of such arrangements is given by $P(n, r)$, defined as:

$$\boxed{P(n,r)=\frac{n!}{(n-r)!}}$$

❒ **Combination** – A combination is an arrangement of r objects from a pool of n objects, where the order does not matter. The number of such arrangements is given by $C(n, r)$, defined as:

$$\boxed{C(n,r)=\frac{P(n,r)}{r!}=\frac{n!}{r!(n-r)!}}$$

Remark: we note that for $0 \leqslant r \leqslant n$, we have $P(n,r) \geqslant C(n,r)$.

Conditional Probability

❒ **Bayes' rule** – For events A and B such that $P(B) > 0$, we have:

$$\boxed{P(A|B)=\frac{P(B|A)P(A)}{P(B)}}$$

Remark: we have $P(A\cap B)=P(A)P(B|A)=P(A|B)P(B)$.

❒ **Partition** – Let $\{A_i, i \in [\![1,n]\!]\}$ be such that for all i, $A_i \neq \varnothing$. We say that $\{A_i\}$ is a partition if we have:

$$\boxed{\forall i \neq j, A_i \cap A_j = \emptyset \quad \textrm{and} \quad \bigcup_{i=1}^{n} A_i = S}$$

Remark: for any event B in the sample space, we have $P(B)=\displaystyle\sum_{i=1}^{n} P(B|A_i)P(A_i)$.

❒ **Extended form of Bayes' rule** – Let $\{A_i, i \in [\![1,n]\!]\}$ be a partition of the sample space. We have:

$$\boxed{P(A_k|B)=\frac{P(B|A_k)P(A_k)}{\displaystyle\sum_{i=1}^{n} P(B|A_i)P(A_i)}}$$

❒ **Independence** – Two events A and B are independent if and only if we have:

$$\boxed{P(A\cap B)=P(A)P(B)}$$

Random Variables

❒ **Random variable** – A random variable, often noted X, is a function that maps every element in a sample space to a real line.

❒ **Cumulative distribution function (CDF)** – The cumulative distribution function F, which is monotonically non-decreasing and is such that $\displaystyle\lim_{x\rightarrow-\infty} F(x)=0$ and $\displaystyle\lim_{x\rightarrow+\infty} F(x)=1$, is defined as:

$$\boxed{F(x)=P(X\leqslant x)}$$

Remark: we have $P(a < X \leqslant B)=F(b)-F(a)$.

❒ **Probability density function (PDF)** – The probability density function f is the probability that X takes on values between two adjacent realizations of the random variable.

❒ **Relationships involving the PDF and CDF** – Here are the important properties to know in the discrete (D) and the continuous (C) cases.

Case	CDF F	PDF f	Properties of PDF
(D)	$F(x)=\displaystyle\sum_{x_i\leqslant x} P(X=x_i)$	$f(x_j)=P(X=x_j)$	$0\leqslant f(x_j)\leqslant 1$ and $\displaystyle\sum_j f(x_j)=1$
(C)	$F(x)=\displaystyle\int_{-\infty}^{x} f(y)dy$	$f(x)=\dfrac{dF}{dx}$	$f(x)\geqslant 0$ and $\displaystyle\int_{-\infty}^{+\infty} f(x)dx=1$

❒ **Variance** – The variance of a random variable, often noted $\textrm{Var}(X)$ or σ^2, is a measure of the spread of its distribution function. It is determined as follows:

$$\boxed{\textrm{Var}(X)=E[(X-E[X])^2]=E[X^2]-E[X]^2}$$

❒ **Standard deviation** – The standard deviation of a random variable, often noted σ, is a measure of the spread of its distribution function which is compatible with the units of the actual random variable. It is determined as follows:

$$\boxed{\sigma=\sqrt{\textrm{Var}(X)}}$$

❒ **Expectation and Moments of the Distribution** – Here are the expressions of the expected value $E[X]$, generalized expected value $E[g(X)]$, k^{th} moment $E[X^k]$ and characteristic function $\psi(\omega)$ for the discrete and continuous cases:

Case	$E[X]$	$E[g(X)]$	$E[X^k]$	$\psi(\omega)$
(D)	$\displaystyle\sum_{i=1}^n x_if(x_i)$	$\displaystyle\sum_{i=1}^n g(x_i)f(x_i)$	$\displaystyle\sum_{i=1}^n x_i^kf(x_i)$	$\displaystyle\sum_{i=1}^n f(x_i)e^{i\omega x_i}$
(C)	$\displaystyle\int_{-\infty}^{+\infty} xf(x)dx$	$\displaystyle\int_{-\infty}^{+\infty} g(x)f(x)dx$	$\displaystyle\int_{-\infty}^{+\infty} x^kf(x)dx$	$\displaystyle\int_{-\infty}^{+\infty} f(x)e^{i\omega x}dx$

Remark: we have $e^{i\omega x} = \cos(\omega x) + i\sin(\omega x)$.

❒ **Revisiting the k^{th} moment** – The k^{th} moment can also be computed with the characteristic function as follows:

$$\boxed{E[X^k] = \frac{1}{i^k}\left[\frac{\partial^k\psi}{\partial\omega^k}\right]_{\omega=0}}$$

❒ **Transformation of random variables** – Let the variables X and Y be linked by some function. By noting f_X and f_Y the distribution function of X and Y respectively, we have:

$$\boxed{f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|}$$

❒ **Leibniz integral rule** – Let g be a function of x and potentially c, and a,b boundaries that may depend on c. We have:

$$\boxed{\frac{\partial}{\partial c}\left(\int_a^b g(x)dx\right) = \frac{\partial b}{\partial c}\cdot g(b) - \frac{\partial a}{\partial c}\cdot g(a) + \int_a^b \frac{\partial g}{\partial c}(x)dx}$$

❒ **Chebyshev's inequality** – Let X be a random variable with expected value μ and standard deviation σ. For $k,\sigma > 0$, we have the following inequality:

$$\boxed{P(|X-\mu| \geqslant k\sigma) \leqslant \frac{1}{k^2}}$$

Jointly Distributed Random Variables

❒ **Conditional density** – The conditional density of X with respect to Y, often noted $f_{X|Y}$, is defined as follows:

$$\boxed{f_{X|Y}(x) = \frac{f_{XY}(x,y)}{f_Y(y)}}$$

❒ **Independence** – Two random variables X and Y are said to be independent if we have:

$$\boxed{f_{XY}(x,y) = f_X(x)f_Y(y)}$$

❒ **Marginal density and cumulative distribution** – From the joint density probability function f_{XY}, we have:

Case	Marginal density	Cumulative function
(D)	$\displaystyle f_X(x_i) = \sum_j f_{XY}(x_i,y_j)$	$\displaystyle F_{XY}(x,y) = \sum_{x_i\leqslant x}\sum_{y_j\leqslant y} f_{XY}(x_i,y_j)$
(C)	$\displaystyle f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dy$	$\displaystyle F_{XY}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(x',y')dx'dy'$

❒ **Distribution of a sum of independent random variables** – Let $Y = X_1 + ... + X_n$ with $X_1,...,X_n$ independent. We have:

$$\boxed{\psi_Y(\omega) = \prod_{k=1}^n \psi_{X_k}(\omega)}$$

❒ **Covariance** – We define the covariance of two random variables X and Y, that we note σ_{XY}^2 or more commonly $\mathrm{Cov}(X,Y)$, as follows:

$$\boxed{\mathrm{Cov}(X,Y) \triangleq \sigma_{XY}^2 = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X\mu_Y}$$

❒ **Correlation** – By noting σ_X, σ_Y the standard deviations of X and Y, we define the correlation between the random variables X and Y, noted ρ_{XY}, as follows:

$$\boxed{\rho_{XY} = \frac{\sigma_{XY}^2}{\sigma_X\sigma_Y}}$$

Remarks: For any X,Y, we have $\rho_{XY} \in [-1,1]$. If X and Y are independent, then $\rho_{XY} = 0$.

❒ **Main distributions** – Here are the main distributions to have in mind:

Type	Distribution	PDF	$\psi(\omega)$	$E[X]$	$\mathrm{Var}(X)$
(D)	$X \sim \mathcal{B}(n,p)$ Binomial	$P(X=x) = \binom{n}{x} p^xq^{n-x}$ $x \in [\![0,n]\!]$	$(pe^{i\omega}+q)^n$	np	npq
	$X \sim \mathrm{Po}(\mu)$ Poisson	$P(X=x) = \frac{\mu^x}{x!}e^{-\mu}$ $x \in \mathbb{N}$	$e^{\mu(e^{i\omega}-1)}$	μ	μ
(C)	$X \sim \mathcal{U}(a,b)$ Uniform	$f(x) = \frac{1}{b-a}$ $x \in [a,b]$	$\frac{e^{i\omega b}-e^{i\omega a}}{(b-a)i\omega}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
	$X \sim \mathcal{N}(\mu,\sigma)$ Gaussian	$f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ $x \in \mathbb{R}$	$e^{i\omega\mu-\frac{1}{2}\omega^2\sigma^2}$	μ	σ^2
	$X \sim \mathrm{Exp}(\lambda)$ Exponential	$f(x) = \lambda e^{-\lambda x}$ $x \in \mathbb{R}_+$	$\frac{1}{1-\frac{i\omega}{\lambda}}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

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.:

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

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:

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

❒ **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:

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

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	$\frac{s^2(n-1)}{\sigma^2}\sim\chi_{n-1}^2$	$\left[\frac{s^2(n-1)}{\chi_2^2},\frac{s^2(n-1)}{\chi_1^2}\right]$

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:

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

❒ ***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	$P(T\leqslant T_0\|H_0\textrm{ true})$	$P(T\geqslant T_0\|H_0\textrm{ true})$	$P(\|T\|\geqslant\|T_0\|\|H_0\textrm{ true})$

❒ **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 $np\geqslant5$
$V\underset{H_0}{\sim}\mathcal{B}\left(n,p=\frac{1}{2}\right)$	$Z=\frac{V-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\underset{H_0}{\sim}\mathcal{N}(0,1)$

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

❒ **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:

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

❒ **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:

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

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

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

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:

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

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

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

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

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

❒ **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:

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

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

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

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

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:

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

❒ **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	$\dfrac{V - \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)}{\frac{1}{\sqrt{n-3}}} \underset{n\gg1}{\sim} \mathcal{N}(0,1)$	$\left[\dfrac{e^{2V_1}-1}{e^{2V_1}+1}, \dfrac{e^{2V_2}-1}{e^{2V_2}+1}\right]$