Chapter 1

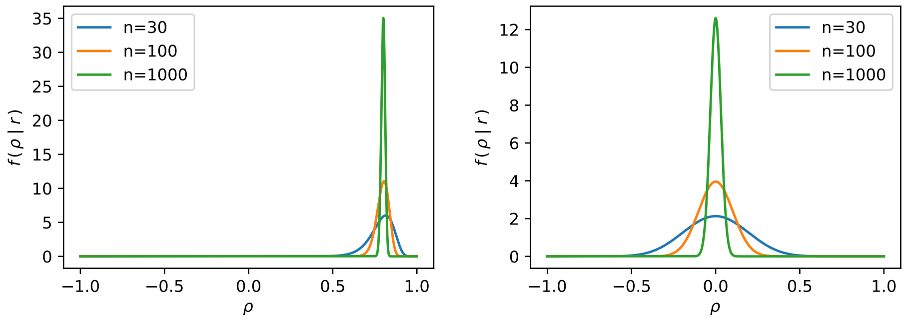
Appendix

1.1 Confidence interval for absolute correlation in bivariate Gaussian

From [?], given a bivariate Gaussian, the confidence distribution of ρ given the empirical correlation r based on n observations is given by

$$f(\rho \mid r, \nu) = \frac{\nu(\nu-1)\Gamma(\nu-1)}{\sqrt{2\pi}\Gamma(\nu+\frac{1}{2})} \frac{\left(1-r^2\right)^{\frac{\nu-1}{2}}\left(1-\rho^2\right)^{\frac{\nu-2}{2}}}{(1-r\rho)^{\frac{2\nu-1}{2}}} F\left(\frac{3}{2}, -\frac{1}{2}, \nu+\frac{1}{2}, \frac{1+r\rho}{2}\right)$$

where $F(a, b, c, z)$ is the Gaussian hypergeometric function and $\nu = n - 1$. That is, given a sample correlation r, what is the confidence in ρ in terms of a distribution. In the following figure, a sample correlation $r = 0.8$ and $r = 0$ has been used with varying number of observations (degrees of freedom) in figures Figure 1.1a and Figure 1.1b respectively. A key property is that f is *even symmetric* in ρ, r. That is $f(\rho \mid r) = f(-\rho \mid -r)$. Thus, a confidence interval for ρ given r is the negative of the confidence interval given $-r$. In particular, if we only observe $|r|$, we can calculate a confidence interval for ρ up to the sign of the bounds of the interval. Furthermore, as we want a CI for $|\rho|$, it does not matter if r is negative or positive. Hence, without loss of generality, we assume that $r \geq 0$. At this point, to construct a confidence interval for $|\rho|$ we



(a) **(b)**

Figure 1.1: $f(\rho \mid r, \nu)$ shown for $r = 0.8$ and $r = 0$ in (a) and (b) with $n \in \{30, 100, 1000\}$. As one would expect, the power i.e. the width of the peak decreases with increasing n and for correlations closer to 0, the width is the largest.

list the following desired properties. Firstly, it should be an exact confidence interval, meaning that for a given significance level α, the CI includes the true value exactly $1-\alpha$ fraction of the times. Secondly, if for a given r, it can not be rejected that ρ is 0, 0 should also be contained in the interval. Finally, if we reject that $\rho = 0$, we shall have $\alpha/2$ probability mass above and below the bounds of the interval. The above is enough to uniquely define a confidence interval in all cases. Before continuing with how this CI is calculated, we mention that as r is an unbiased estimator of ρ, we would preferably want $|r| \in CI_{1-\alpha}(|\rho|)$ (where $CI_{1-\alpha}(|\rho|)$ denotes the $1-\alpha$ confidence interval for $|\rho|$). However, although this will in almost every scenario be the case, we can not be sure of this from the above properties and in fact examples with large α can be constructed such that $|r|$ lies just outside the constructed CI.

First, to conform with the second desired property, if it can not be rejected that $\rho = 0$ on a significance level α, we will initially compute a CI for ρ (not $|\rho|$) based on r (wlog chosen to be non-negative). This CI will just be a symmetric CI in the sense that $\alpha/2$ of the probability mass will lie below the lower bound of the CI and above the upper bound of the CI respectively. If 0 is contained in this CI, we can not reject that $\rho = 0$ and vice versa on an α significance level. Thus, if 0 is contained in this initial CI for ρ, we will start the CI for $|\rho|$ at 0 and determine and upper bound b such that α probability mass is above this b. Otherwise, we shall find a and b such that $\alpha/2$ probability mass is below a and above b respectively. Choosing a and b this way also conforms with the third property. Finally, to ensure that the CI contains exactly $1-\alpha$, we define $\tilde{f}$ as the reflected f in ρ such that

$$\tilde{f}\,(\rho_a \mid r_a, \nu) = f(\rho_a \mid r_a, \nu) + f(-\rho_a \mid r_a, \nu), \quad \rho_a, r \in [0, 1]$$

where ρ_a and r_a is the absolute correlation and empirical correlation respectively. With this $\tilde{f}$, the density at ρ_a is both the density for the negative and positive correlation ensuring that the $\tilde{f}$ has probability mass 1. Thus, if $a = 0$ (i.e. the CI must contain 0), we find b as the $1 - \alpha$ percentile of $\tilde{f}$ and if $a \neq 0$, we take a as the $\alpha/2$ percentile and b as the $1 - \alpha/2$ percentile of $\tilde{f}$.

As an example, suppose $r_a = 0.06$ with 1000 observations. Then a 95% CI for $|\rho|$ is $[0, 0.11164]$ whereas if on had observed $r_a = 0.07$ the CI would be $[0.01071, 0.1314]$. These CI could then be used to test the absolute correlation of a bivariate Gaussian i.e. for $r_a = 0.07$ based on 1000 observations would be rejected as stemming from a Gaussian with absolute correlation 0.01 on a 5% significance level.