Misc. Maths for Microeconometrics

Contents

Disclaimer: sections and lines in brown are 'under construction'.

1 Probability Theory

1.1 Definitions

1.1.1 Random variables and probability distributions

Random variables

A **random variable X** is a numerical representation of the outcome of a random process. Formally, it's a function $X\colon S \mapsto E$ that maps each process outcome in a sample space S to a measurable space E. It may be discrete, continuous, or mixed. Examples:

- discrete: the outcome of a coin toss: $\{X(\text{heads}) = 1, X(\text{tails}) = 0\}$;
- continuous: $X \sim \mathcal{U}[a, b]$.

The **probability distribution** of X is the probability measure on E defined by $\mathrm{P}_X(A) \coloneqq \Pr(X \in A)$ for every measurable subset $A \subseteq E$. When $E \subseteq \mathbb{R}$, it may be described by:

- a **probability mass or density function f_X**:
  - for a discrete X, $f_X(x) = \Pr(X = x)$ is its probability mass function;
  - for a continuous X, $f_X(x)$ is its probability density function, so that $\Pr(X \in A) = \int_A f_X(t)\,dt$.
- its **cumulative distribution function** (CDF)

$$F_X(x) \coloneqq \Pr(X \le x) \;=\; \begin{cases} \sum_{t \le x} f_X(t) & \text{for a discrete } X \\ \int_{-\infty}^{x} f_X(t)\,dt & \text{for a continuous } X \end{cases}$$

  The **α-quantile** of X is the smallest value of X below which lies $100\alpha\%$ of its distribution, and is defined by the CDF's generalized inverse

$$\mathbb{Q}_X(\alpha) \coloneqq F_X^{-1}(\alpha) \coloneqq \inf\{x \in \mathbb{R} : F_X(x) \ge \alpha\}, \qquad \alpha \in (0, 1).$$

We generally write $X \sim \mathcal{D}$ when X follows a named distribution $\mathcal{D}$, or say that X has distribution P_X, density f_X, or CDF F_X; and may also informally write $X \sim f_X$.

A group of random variables $(Z_1, Z_2, \ldots)$ has a **joint probability distribution,** which describes how the variables vary together. When it admits a probability mass or density function, its joint function $f_{Z_1, Z_2, \ldots}$ can be factorized into marginal and conditional functions using the chain rule. Example:

$$f_{X,Y}(x, y) \;=\; f_X(x)\, f_{Y|X}(y|x) \;=\; f_Y(y)\, f_{X|Y}(x|y)$$

1.1.2 Every probability distribution has its moments...

We often focus on the first few moments—including the central, standardized, and mixed moments—of probability distributions:[1]

- For an individual random variable X:

[1] The r^{th} moment of X's probability distribution is $\mathbb{E}[X^r]$. Its r^{th} *central* moment is $\mathbb{E}\left[(X - \mathbb{E}[X])^r\right]$. Its r^{th} *standardized* moment is $\mathbb{E}\left[(X - \mathbb{E}[X])^r\right] / \mathbb{V}[X]^r$.

Expectation $\mathbb{E}[X] := \int x f_X(x)\,dx = \int x\,dF_X(x)$ is the distribution's center of mass or mean. It is often written μ_X.

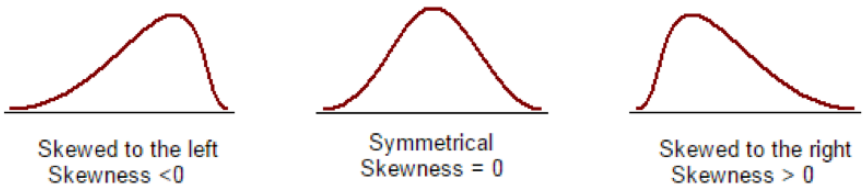
Variance $\mathbb{V}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ is a measure of dispersion, i.e., how far values are spread out from their average. It is often written σ_X^2.

Standard deviation $\mathrm{SD}[X] := \sqrt{\mathbb{V}[X]} = \sigma_X$ is the square root of the variance, and thereby also a measure of spread. It is more easily interpretable than the variance as it is expressed in the same unit as X.

Skewness $\mathrm{Skew}[X] := \dfrac{\mathbb{E}[(X - \mathbb{E}[X])^3]}{\sigma_X^3}$ is a measure of asymmetry.[2]

For a unimodal continuous X:

- $\mathrm{Skew}[X] < 0$ means a "skew to the left", aka a left tail
- $\mathrm{Skew}[X] > 0$ means a "skew to the right", aka a right tail



- For a pair of random variables (X, Y):

Covariance $\mathrm{cov}[X, Y] := \mathbb{E}\big[\big(X - \mathbb{E}[X]\big)\big(Y - \mathbb{E}[Y]\big)\big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ is a measure of the joint variability of two random variables. It is often written $\sigma_{X,Y}^2$. Because it changes with the units of the random variables proportionally, it is not conducive to comparisons of the relationship strength between different pairs of random variables that do not have the same units.

Correlation $\rho_{X,Y} := \dfrac{\mathrm{cov}[X, Y]}{\sigma_X \sigma_Y} \quad \in [-1, 1]$

Pearson's correlation coefficient is the normalized covariance. As it is unitless, it is conducive to comparisons between different pairs of random variables.

⚠ Covariance and correlation are measures of *linear* relationship only. $\rho_{X,Y}$ is the slope of the regression of standardized Y on standardized X, i.e., the strength of the *linear* relation. A correlation of 0 only indicates that the variables have no *linear* relationship, while they may have a nonlinear relationship—that one may observe using a scatterplot.

1.1.3 ... of which we can only estimate sample equivalents

In statistical analysis, we generally adopt a view of the world in three levels:

(1) the generic-unit population level or DGP, where Y and X are generic random variables;

(2) the random-sample level, which is a sample from that population, i.e., a finite set of random variables generated by the DGP, such as $\{(Y_i, X_i)\}_{i=1}^n$;

(3) the realized-sample level, i.e., the variables' realized values, such as $\{(y_i, x_i)\}_{i=1}^n$.

From a random sample of variables $\{(Y_i, X_i)\}_{i=1}^n$ generated from the DGP, we can define sample equivalents or 'estimators' of the moments of the generic X and Y's individual and joint probability distributions. These quantities are functions of random variables, and thus random variables themselves. We can then use the variable realizations—if observed—to compute estimated values. The table below shows these estimators, as

[2] ⚠ A symmetric distribution has $Skew = 0$, but the reverse isn't true. E.g., a distribution with one long but thin tail, and one short but fat tail, will have $Skew = 0$.

well as a correction for bias when applicable:[3]

Population parameter	Sample equivalent	Bias-adjusted sample equivalent
$\mu_X = \mathbb{E}[X]$	$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$	$\bar{X} := \frac{1}{n}\sum_i X_i$
$\sigma_X^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$	$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$	$\frac{n}{n-1}\times\ s_X^2 := \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$
$\sigma_{X,Y}^2 = \mathbb{E}\big[\big(X - \mathbb{E}[X]\big)\big(Y - \mathbb{E}[Y]\big)\big]$	$s_{X,Y}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$	$\frac{n}{n-1}\times\ s_{X,Y}^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})(Y_i - \bar{Y})$

1.1.4 Robust measures: *median, interquantile range...*

The mean and the standard deviation, as measures of centrality and dispersion, respectively, are heavily influenced by the magnitude of values. They can be misleading with skewed data, and are very sensitive to outliers. We can use alternative measures that are rank-based and thereby more robust to outliers, such as:

- as measures of central tendency: the **median $\mathbf{med[X] := \mathbb{Q}_X(0.5)}$**.
- as measures of statistical dispersion:
  - The **interquantile range $\mathbf{IQR[X] := \mathbb{Q}_X(.75) - \mathbb{Q}_X(.25)}$** which is the difference between the upper and lower quartiles.
  - The **median absolute deviation $\mathbf{MAD[X] := med\left[|X - med[X]|\right]}$** which is the median of the absolute deviations from the median.

1.1.5 Standardizing: *z-score*

A standardized variable or "z-score" is a variable that has been centered and scaled to have mean 0 and standard deviation 1. The standardized variable measures how many standard deviations the raw variable is from the mean.

$$\text{Population: } Z = \frac{X - \mu_X}{\sigma_X}, \qquad \text{Sample: } Z = \frac{X - \bar{X}}{s_X}$$

1.1.6 Independence ⇒ uncorrelation = orthogonality

Independence and correlation are statistical concepts, whereas orthogonality is a linear algebra concept.

- Two random variables X and Y are **independent ($\boldsymbol{X \perp\!\!\!\perp Y}$)** iff $f_{X,Y}(x,y) = f_X(x) f_Y(y),\ \forall (x,y)$.
- Two random variables X and Y are **uncorrelated** iff $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, i.e., $\text{cov}[X,Y] = 0$.
- Two vectors u and v are **orthogonal** iff their inner product $<u, v> = 0$
  A space of random variables can be considered a vector space. We can therefore define an inner product in that space, in various ways. One option is to define it as the covariance: $<X, Y> := \text{cov}[X,Y]$. Two random variables X and Y are therefore **orthogonal** iff $\text{cov}[X,Y] = 0$.

$$\begin{array}{ccc} \text{independent} & \Longrightarrow & \text{uncorrelated, orthogonal} \\ X \perp\!\!\!\perp Y & & \text{cov}[X,Y] = 0 \end{array}$$

[3] The bias of an estimator $\hat{\theta}$ for an estimand θ is $\mathbb{E}[\hat{\theta} - \theta]$; an *unbiased* estimator of θ is a random variable $\hat{\theta}$ s.t. $\mathbb{E}[\hat{\theta}] = \theta$. In the formulas of the sample variance and covariance, a bias occurs because we lost one degree of freedom in calculating the sample mean $\bar{X}$ that the formulas use in place of the population mean $\mathbb{E}[X]$. For example, in the case of the variance, we can show that $\mathbb{E}[s_X^2] = ... = \frac{n-1}{n}\sigma_X^2 \neq \sigma_X^2$. So we must correct by a factor of $\frac{n}{n-1}$ to recover an unbiased estimator of σ^2.

1.2 Common families of probability distributions

1.2.1 Discrete

Many discrete probability distributions are built from the concept of Bernouilli trials. A Bernouilli(p) trial is a random experiment where the probability of success—which we can equivalently view as the occurrence of an event—is p. The number of occurrences of an event in a sequence of n observations is thereby equivalent to the number of successes in a sequence of n *iid* Bernoulli(p) trials. In the table below, we hence "event" and "success" interchangeably.

Name	**Description & Support**	**PDF** $\mathbf{Pr(X{=}x \mid \theta) = \ldots}$	**Moments** $\mathbb{E}[X\|\theta],\ \mathbb{V}[X\|\theta]$
Bernoulli $X \sim \text{Ber}(p)$	X is the outcome of a single Bernouilli trial with probability p of success: $X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$	$= p^x(1-p)^{1-x}$	$\mathbb{E} = p$ $\mathbb{V} = p(1-p)$
Binomial $X \sim \text{binomial}(n, p)$	X is the number of successes in n independent Bernoulli(p) trials. $X \in \{0, 1, ..., n\}$	$= \binom{n}{x} p^x(1-p)^{n-x}$	$\mathbb{E} = np$ $\mathbb{V} = np(1-p)$
Poisson $X \sim \text{Pois}(\lambda)$	X is the number of events occurring in a fixed interval, assuming events occur independently at a constant average rate λ. *There is no finite upper bound.* $X \in \{0, 1, ...\}$	$= \dfrac{\lambda^x e^{-\lambda}}{x!}$	$\mathbb{E} = \lambda$ $\mathbb{V} = \lambda$
Generalized Poisson $X \sim \text{GPois}(\lambda, \eta)$	Same setting as for the Poisson distribution, with an additional dispersion parameter $0 \leq \eta < 1$. $X \in \{0, 1, ...\}$	$= \frac{\lambda(\lambda+\eta x)^{x-1}}{x!} e^{-\lambda-\eta x}$	$\mathbb{E} = \frac{\lambda}{1-\eta}$ $\mathbb{V} = \frac{\lambda}{(1-\eta)^3}$
Negative Binomial $X \sim \text{NB}(r, p)$	X is the number of successes in a sequence of *iid* Bernoulli(p) trials before a number r of failures occurs. $\frac{1}{r}$ is called the dispersion parameter. $X \in \{0, 1, ...\}$	$= \binom{x+r-1}{x} p^x(1-p)^r$	$\mathbb{E} = \frac{pr}{1-p}$ $\mathbb{V} = \frac{pr}{(1-p)^2} = \mathbb{E} + \frac{\mathbb{E}^2}{r}$

$\diamond$ *Note on the count distributions with unbounded support.* The Poisson distribution is equidispersed: $\mathbb{V}[X] = \mathbb{E}[X]$. This assumption is often unrealistic as count data are usually overdispersed, i.e., their variance is higher than their mean. A Poisson regression model will hence not be able to adequately explain the variability in the observed counts, and should have poor prediction ability. With overdispersed count data, one may instead choose the negative binomial or the generalized Poisson distributions, which have $\mathbb{V}[X] > \mathbb{E}[X]$. As the dispersion parameter gets smaller, the variance converges to the mean, and the given distribution converges to a Poisson distribution.

1.2.2 Continuous

Name	Description & Support	PDF $\Pr(X{=}x \mid \theta) = \ldots$	Moments $\mathbb{E}[X\|\theta],\ \mathbb{V}[X\|\theta]$
Uniform $X \sim \mathcal{U}(a,b)$	X is the outcome from a trial that is limited between two bounds: $X \in [a,b]$	$= \frac{1}{b-a}$	$\mathbb{E} = \frac{1}{2}(a+b)$ $\mathbb{V} = \frac{1}{12}(b-a)^2$
Beta $X \sim \text{Beta}(\alpha,\beta)$	X is a process limited to intervals of finite length, such as a percentage or a proportion.[4] $\alpha, \beta > 0$ are two shape parameters. $X \in [0,1]$	$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\mathbb{E} = \frac{\alpha}{\alpha+\beta}$ $\mathbb{V} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Gamma[5] $X \sim \text{Gamma}(\alpha,\beta)$	$\alpha > 0$ is a shape parameter and $\beta > 0$ a rate or 'inverse scale' parameter. $X \in (0,\infty)$	$= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\mathbb{E} = \frac{\alpha}{\beta}$ $\mathbb{V} = \frac{\alpha}{\beta^2}$
Exponential $X \sim \text{Exp}(\beta)$	X is the distance between events in a Poisson point process (= a process in which events occur continuously and independently at a constant average rate). It is a special case of the gamma distribution: $\Gamma(1,\beta)$. $X \in [0,\infty)$	$= \beta e^{-\beta x}$	$\mathbb{E} = \frac{1}{\beta}$ $\mathbb{V} = \frac{1}{\beta^2}$
Chi-squared $X \sim \chi^2_k$	X is a sum of the squares of k independent standard normal random variables. The χ^2 distribution is used primarily in hypothesis testing. It is a special case of the gamma distribution: $\Gamma\left(\frac{k}{2}, \frac{1}{2}\right)$. $X \in (0,\infty)$	$= \frac{2^{-\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$	$\mathbb{E} = k$ $\mathbb{V} = 2k$
Logistic $X \sim \text{Logistic}(\mu,s)$	X's cumulative distribution function is the logistic function (which appears in logistic regression). Its shape resembles the normal distribution's but has heavier tails (higher kurtosis). $X \in (-\infty,\infty)$	$= \frac{e^{\frac{\mu-x}{s}}}{s\left(1+e^{\frac{\mu-x}{s}}\right)^2}$	$\mathbb{E} = \mu$ $\mathbb{V} = \frac{s^2\pi^2}{3}$
Normal $X \sim \mathcal{N}(\mu,\sigma^2)$	$X \in (-\infty,\infty)$	$= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$\mathbb{E} = \mu$ $\mathbb{V} = \sigma^2$
Student's $\mathcal{T}$ $X \sim \mathcal{T}_k$	$X \in (-\infty,\infty)$	$= \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)} \left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}$	$\mathbb{E} = 0$ for $k > 1$ $\mathbb{V} = \frac{k}{k-2}$ for $k > 2$

The normal distribution is ubiquitous in statistics because many statistics—such as sample means, differences in means, and regression coefficient estimators—can be expressed as weighted averages or sums of many observations, whose sampling distributions Central Limit Theorems imply converge to a normal distribution as the sample size n increases.

[4] Say we want a bell-shape distribution, defined by its mean and standard deviation, but that needs to be bounded to a given interval — such that the normal distribution is not appropriate. Ex: test scores data are bounded to [0-100]. Should we use a Beta distribution or a truncated normal? A. Gelman's recommendation: use a truncated normal, i.e., use a normal, and if we get some simulated data at 104, transform them to 100. This is sort of representing the underlying process: individuals whose ability would truly take them above 100, but the test isn't able to account for such ability levels, so they get 100. Instead, the Beta distribution is very abstract and does not represent any type of underlying process.

[5] The Gamma distribution is a common choice to model right-skewed continuous data. It also has a specific mean-variance relationship: the variance of the data increases with the square of the mean. It can therefore be a suitable distribution for data whose standard deviation might be approximately proportional to the mean.

1.2.3 Linear Exponential Families

A probability density function is said to be of a linear exponential family (LEF) iff it can be expressed as

$$f_X(x) = \exp(a(\theta) + b(x) + c(\theta)x)$$

where $b(x)$ is a normalizing constant that ensures probabilities sum or integrate to one.

All generalized linear models are based on an assumed LEF density. Such densities have $\mathbb{E}[X] = -c(\theta)^{-1}a'(\theta)$ and $\mathbb{V}[X] = c(\theta)^{-1}$.

Examples include:[6]

Name	**Density $f_X(x)$ as $\exp(a(\theta) + b(x) + c(\theta)x)$**	$\mathbb{E}[X\|\theta]$	$\mathbb{V}[X\|\theta]$
Normal $X \sim \mathcal{N}(\mu, \sigma^2)$	$= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \ldots = \exp\left(\frac{-\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) - \frac{y^2}{2\sigma^2} + \frac{\mu}{\sigma^2}y\right)$	$\mathbb{E} = \mu$	$\mathbb{V} = \sigma^2$
Bernoulli $X \sim \text{Ber}(p)$	$= p^x(1-p)^{1-x} = \ldots = \exp\left(\ln(1-p) + \ln\left(\frac{p}{1-p}\right)x\right)$	$\mathbb{E} = p$	$\mathbb{V} = p(1-p)$
Exponential $X \sim \text{Exp}(\beta)$	$= \beta e^{-\beta x} = \ldots = \exp(\ln\beta - \beta x)$	$\mathbb{E} = \frac{1}{\beta}$	$\mathbb{V} = \frac{1}{\beta^2}$
Poisson $X \sim \text{Pois}(\lambda)$	$= \frac{\lambda^x e^{-\lambda}}{x!} = \ldots = \exp(-\lambda - \ln x! + x\ln\lambda)$	$\mathbb{E} = \lambda$	$\mathbb{V} = \lambda$

1.2.4 Conjugate prior probability distributions in Bayesian inference

Let Y denote the random outcome variable in a regression model, θ an unknown population parameter of interest, represented by the random variable Θ. For convenience, we drop the conditioning on the regressors X in the subsequent formulas.
In Bayesian inference, if the prior and posterior probability distributions $f_\Theta(\theta)$ and $f_{\Theta|Y}(\theta|y)$ are in the same family, this family is called a *conjugate prior distribution* for the distribution that is the likelihood function $f_{Y|\Theta}(y|\theta)$, i.e., for the sampling model. Having a conjugate prior is convenient, as it gives an algebraic expression for the posterior. For example:

- The Beta distribution, for $\Theta \in [0,1]$, is a conjugate prior for the Bernoulli, binomial, negative binomial and geometric distributions.

  Ex:

  $$\begin{cases} \text{sampling model:} & Y_i|\Theta \sim \text{binomial}(n,p) \\ \text{prior:} & \Theta \sim \text{Beta}(a,b) \end{cases} \implies \text{posterior:} \quad \Theta \mid Y_i = y_i \sim \text{Beta}(a + y_i, b + n - y_i)$$

  - Generalization: The Dirichlet distribution, for a vector of probabilities that must sum to 1, is a conjugate prior for the categorical and multinomial distributions;

- The Gamma distribution, for a rate (inverse scale) parameter, is a conjugate prior for a Poisson or exponential distribution.

  Ex:

  $$\begin{cases} \text{sampling model and prior:} \\ Y_i|\Theta \sim \text{Poisson}(\Theta) \\ \Theta \sim \text{Gamma}(a,b) \end{cases} \implies \begin{cases} \text{posterior predictive distribution and posterior:} \\ \widetilde{Y}_i \mid Y_i = y_i \sim \text{NB}\left(a + \sum y_i, b + n\right) \\ \theta \mid Y_i = y_i \sim \text{Gamma}\left(a + \sum y_i, b + n\right) \end{cases}$$

- The Gamma distribution, for a precision (inverse variance) parameter, is a conjugate prior for a normal distribution with known mean.

[6] But also the binomial with number of trials known (of which the Bernoulli is a special case), some negative binomial models (of which the geometric and the Poisson are special cases), and the one-parameter gamma (of which the exponential is a special case).

Ex:

$$\begin{cases} \text{sampling model and priors:} \\ Y_i \mid \Theta, \sigma^2 \sim \mathcal{N}(\Theta, \sigma^2) \\ \Theta \mid \sigma^2 \sim \mathcal{N}\left(\mu_0, \dfrac{\sigma^2}{\kappa_0}\right) \\ \dfrac{1}{\sigma^2} \sim \text{Gamma}\left(\dfrac{\nu_0}{2}, \dfrac{\nu_0 \sigma_0^2}{2}\right) \end{cases} \Longrightarrow \begin{cases} \text{posteriors:} \\ \Theta \mid Y_i, \sigma^2 \ \sim \mathcal{N}\left(\mu_n, \dfrac{\sigma^2}{\kappa_n}\right) \\ \dfrac{1}{\sigma^2} \mid Y_i \ \sim \text{Gamma}\left(\dfrac{\nu_n}{2}, \dfrac{\nu_n \sigma_n^2}{2}\right) \end{cases}$$

– Generalization: The Wishart distribution, for a symmetric non-negative definite matrix, is a conjugate prior for a multivariate normal distribution (specifically: for the inverse of its covariance matrix).

1.3 Convergence theorems

Many estimators and test statistics are made of sample averages. We are therefore interested in how sequences of sample averages behave as a sample size $n \to \infty$. Two groups of theorems come into play:

- **Laws of Large Numbers (LLNs)** say that sample averages converge in probability;
- **Central Limit Theorems (CLTs)** say that sample averages converge in distribution.

Convergence

- Consider a sequence of real numbers, $(a_n)_{n\in\mathbb{N}^*} := (a_1, a_2, \ldots)$. This deterministic sequence may converge, in which case to a limit a_∞, and we write $a_n \xrightarrow[n\to\infty]{} a_\infty$, or equivalently, $\lim_{n\to\infty} a_n = a_\infty$.
  - *For example, let* $a_n = 4 + \frac{5}{n}$. *We have* $a_n \xrightarrow[n\to\infty]{} 4$.
- Consider a sequence of random variables, $(X_n)_{n\in\mathbb{N}^*} := (X_1, X_2, \ldots)$. This stochastic sequence may converge, in which case to a random variable X_∞, always under a specific mode of stochastic convergence. For example:
  - Convergence in probability:[a] $X_n \xrightarrow[n\to\infty]{p} X \iff \operatorname{plim}_{n\to\infty} X_n = X$;
  - Convergence in distribution (weaker): $X_n \xrightarrow[n\to\infty]{d} X \iff \lim_{n\to\infty} F_{X_n} = F_X$.

[a] Formally, a series converges iff it will eventually be within any small distance ε of its limit:
- a_n converges to a_∞ iff $\forall \varepsilon > 0, \exists\, m$ s.t. $\forall n \geqslant m, |a_n - a_\infty| < \varepsilon$.
- X_n converges *in probability* to X iff $\forall \varepsilon > 0$, $\Pr(|X_n - X| > \varepsilon) \xrightarrow[n\to\infty]{} 0$.

Consider a sample of n random variables $\{X_i\}_{i=1}^n$, and their sample average $\bar{X}_N := \frac{1}{N}(X_1 + \ldots + X_N)$. We are interested in the convergence of this (sequence of) sample averages $\left(\bar{X}_n\right)_{n\in\mathbb{N}^*} := \left(\bar{X}_{n_1}, \bar{X}_{n_2}, \ldots\right)$, succinctly noted $\bar{X}_n$. We distinguish three situations based on the underlying sample:[7]

a. The $\{X_i\}$ are **not independent** over i

b. The $\{X_i\}$ are **independent** over i but not identically distributed: $X_i \sim (\mu_i, \sigma_i^2), \forall i$

c. The $\{X_i\}$ are **independent and identically distributed (iid)**: $X_i \overset{\text{iid}}{\sim} (\mu, \sigma^2), \forall i$

Laws of Large Numbers (LLNs)

a. The average $\bar{X}_n$ of n random variables converges in probability:

$$\bar{X}_n - \mathbb{E}[\bar{X}_n] \xrightarrow[n\to\infty]{p} 0 \iff \operatorname{plim}_{n\to\infty} \bar{X}_n = \lim_{n\to\infty} \mathbb{E}[\bar{X}_n]$$

b. If the X_i are independent, the probability limit simplifies to: $\operatorname{plim}_{n\to\infty} \bar{X}_n = \lim_{n\to\infty} \frac{1}{n} \sum_i \mathbb{E}[X_i]$

c. If the X_i are iid: $\operatorname{plim}_{n\to\infty} \bar{X}_n = \lim_{n\to\infty} \frac{1}{n} n\mu = \mu$

Central Limit Theorems (CLTs)

a. The *standardized* average[a] of n random variables converges to a normal distribution, *regardless*

[7] This section draws heavily from Colin Cameron's lecture notes "Asymptotic Theory for OLS".

of the distribution of the original variables:

$$\frac{\bar{X}_n - \mathbb{E}[\bar{X}_n]}{\sqrt{\mathbb{V}[\bar{X}_n]}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

b. If the X_i are independent:

$$\frac{\frac{1}{n}\sum_i \left(X_i - \mathbb{E}[X_i]\right)}{\frac{1}{n}\sqrt{\sum_i \mathbb{V}[X_i]}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

c. If the X_i are iid:

$$\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

To express results in terms of $\bar{X}_n$, we say that $\bar{X}_n$ is *asymptotically normally distributed*:[b]

a. $\bar{X}_n \overset{\text{a}}{\sim} \mathcal{N}\left(\lim_{n\to\infty} \mathbb{E}[\bar{X}_n],\ \lim_{n\to\infty} \mathbb{V}[\bar{X}_n]\right)$

b. $\bar{X}_n \overset{\text{a}}{\sim} \mathcal{N}\left(\lim_{n\to\infty} \frac{1}{n}\sum_i \mathbb{E}[X_i],\ \lim_{n\to\infty} \frac{1}{n^2}\sum_i \mathbb{V}[X_i]\right)$

c. $\bar{X}_n \overset{\text{a}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

[a] By an LLN, $\bar{X}_n$ has a degenerate distribution as it converges to a constant $\mathbb{E}[\bar{X}_n]$. To apply the CLT, we first standardize it $\bar{X}_n$—using its expectation and standard deviation—to construct a random variable with variance 1, i.e., with a nondegenerate distribution.

[b] The asymptotics here correspond to a n is large enough that it's reasonable to consider the approximation, but not so large that the asymptotic variance goes to zero and makes the distribution degenerate.

Examples:

- Let X_i denote the result of a coin flip, with $X_i \coloneqq 1$ for heads and 0 for tails. We consider a sample of n i.i.d. coin flips, with sample average $X_n \coloneqq \frac{1}{n}(X_1 + ... + X_n)$ which is the proportion of heads.
  - By the LLN, $\bar{X}_n \xrightarrow[n\to\infty]{p} \frac{1}{2}$.
  - By the CLT, $\bar{X}_n \overset{\text{a}}{\sim} \mathcal{N}\left(\frac{1}{2}, ...,\right)$, i.e., the probability of getting $\frac{n}{2}$ heads will get increasingly close to a normal distribution as n increases.
- LLNs and CLTs are widely used in econometrics because extremum estimators can be written as averages. $\hat{\beta}_{\text{OLS}}$ is the statistic whose convergence we are interested in.
  - LLNs give consistency. Ex: we can rewrite $\hat{\beta}_{\text{OLS}}$ to make two averages appear and apply LLNs:

$$\hat{\beta}_{\text{OLS}} = \beta_0 + (\mathsf{X}'\mathsf{X})^{-1}\mathsf{X}'u = \beta_0 + \left(\frac{1}{n}\mathsf{X}'\mathsf{X}\right)^{-1}\frac{1}{n}\mathsf{X}'u = \beta_0 + \underbrace{\left(\frac{1}{n}\sum_i \mathrm{X}_i\mathrm{X}_i'\right)^{-1}}_{\xrightarrow[n\to\infty]{p} M_{\mathsf{XX}}} \underbrace{\frac{1}{n}\sum_i \mathrm{X}_i u_i}_{\xrightarrow[n\to\infty]{p} \mathbb{E}[\mathrm{X}_i u_i] = 0} \xrightarrow[n\to\infty]{p} \beta_0$$

  - CLTs give asymptotic distributions. Ex: we can center and rescale $\hat{\beta}_{\text{OLS}}$ to apply a CLT:

$$\sqrt{n}\left(\hat{\beta}_{\text{OLS}} - \beta\right) = \left(\frac{1}{n}\mathsf{X}'\mathsf{X}\right)^{-1}\frac{1}{\sqrt{n}}\mathsf{X}'\mathrm{u} = \underbrace{\left(\frac{1}{n}\sum_i \mathrm{X}_i\mathrm{X}_i'\right)^{-1}}_{\xrightarrow[n\to\infty]{p} M_{\mathsf{XX}}} \underbrace{\frac{1}{\sqrt{n}}\sum_i \mathrm{X}_i u_i}_{\xrightarrow[n\to\infty]{d} \mathcal{N}(0, M_{\mathsf{X}\Omega\mathsf{X}})} \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, M_{\mathsf{XX}}^{-1} M_{\mathsf{X}\Omega\mathsf{X}} M_{\mathsf{XX}}^{-1}\right)$$

2 Linear algebra

2.1 Invertible and positive-definite matrices

Invertibility and positive definiteness are central in econometrics. These matrix properties ensure that linear systems have unique solutions, as required for example for the solution of the OLS estimator of the classical linear regression model $(\mathsf{X}'\mathsf{X})^{-1}\mathsf{X}'\mathsf{Y}$.

Invertibility

A square $k \times k$ matrix A is **invertible** if there exists a $k \times k$ matrix B such that $\mathsf{AB} = \mathsf{BA} = \mathsf{I}_k$. Otherwise, it is called **singular or degenerate**.

The quasi-totality of square matrices are invertible.

Positive definiteness

Let M be a $k \times k$ symmetric real matrix with eigenvalues $\{\lambda_k\}$.

- M is **positive semidefinite** $\iff$ $\mathrm{z}'\mathsf{M}\mathrm{z} \geq 0$ for every vector $\mathrm{z} \in \mathbb{R}^k$ $\iff$ all $\{\lambda_k\}$ are ≥ 0.
- M is **positive definite** $\iff$ $\mathrm{z}'\mathsf{M}\mathrm{z} > 0$ for every vector $\mathrm{z} \in \mathbb{R}^k$ $\iff$ all $\{\lambda_k\}$ are > 0.

Examples:

- The identity matrix I_k is positive definite.
- Population covariance matrices are positive semidefinite—and positive definite unless some variables are exact linear functions of the others. Conversely, every positive-semidefinite matrix is the covariance matrix of some multivariate distribution.

Note: If M *is positive definite, then it is invertible—and* M^{-1} *also.*

3 Calculus and transformations

3.1 Log transformations in regressions

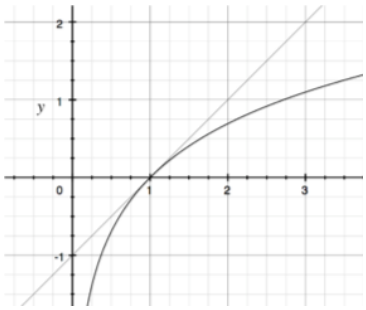
3.1.1 Interpreting coefficients

Regression model		*For a given difference in X, what difference do we expect in the conditional mean of Y?*
level-level (linear)	$Y = \beta_0 + \beta_1 X + u$	A 1-unit difference in X is associated with a β_1-unit average difference in Y.
log-level (log-linear)	$\ln(Y) = \beta_0 + \beta_1 X + u$	A 1-unit difference in X is associated with: – a β_1-unit average difference in $\ln(Y)$ – $\Leftrightarrow$ a e^{β_1}-fold average difference in Y – $\Leftrightarrow$ a $(100(e^{\beta_1}-1))\%$ average difference in Y – *if $\hat{\beta}$ is small, we can approximate $e^{\beta_1}-1 \approx \beta_1$, hence : a $\approx (100\beta_1)\%$ average difference in Y*[*]
level-log	$Y = \beta_0 + \beta_1 \ln(X) + u$	A 1-unit difference in $\ln(X)$ is associated with a β_1-unit average difference in Y. $\Leftrightarrow$ A 1% difference in X is associated with a $\approx \frac{\beta_1}{100}$-unit average difference in Y.
log-log	$\ln(Y) = \beta_0 + \beta_1 \ln(X) + u$	A 1% difference in X is associated with a $\approx \beta_1\%$ average difference in Y. *(β_1 is the elasticity of Y w.r.t. X.)*

[*]Small natural-log differences can be interpreted as proportional changes

The log function is approximately linear around 1. It is therefore reasonable to do first-order Taylor approximation of ln() around 1:

$$\begin{aligned} \text{Around } z = 1, \qquad f(z) &\simeq f'(1)(z-1) + f(1) \\ \ln(z) &\simeq \tfrac{1}{1}(z-1) + 0 \\ \ln(z) &\simeq z - 1 \end{aligned}$$



Thus, for y_2 close to y_1, $\ln(y_2) - \ln(y_1) = \ln\left(\frac{y_2}{y_1}\right) \simeq \frac{y_2}{y_1} - 1 = \frac{y_2 - y_1}{y_1}$

Hence, a small difference in the logs of Y, multiplied by 100, approximately equals the percentage difference in Y. More generally, the exact percentage difference corresponding to a log difference $\Delta \ln(Y)$ is $100\left(e^{\Delta \ln(Y)} - 1\right)\%$.

Therefore, in the log-level regression $\ln(Y) = \beta_0 + \beta_1 X + u$:

- if $|\hat{\beta}_1|$ small, one may say *"a 1-unit difference in X is associated with a $\approx 100\hat{\beta}_1\%$ difference in Y"*;
- if $|\hat{\beta}_1|$ large, one may only say *"a 1-unit difference in X is associated with a $e^{\hat{\beta}_1}$-fold difference in Y"*.

3.1.2 Log vs IHS transformations of the outcome

A common approach for modeling a right-skewed outcome on the positive real line is to log-transform it, $\ln(Y)$.[8] Indeed, the log transformation effectively pulls in high Y values, brings them closer to the middle, narrowing the distribution's range, such that they don't have such a large effect on the results and the outcome is normal enough that linear regression is valid. The assumption of the linear regression model is then that the *logged* outcome is conditionally normally distributed: $\ln(Y|X_1, ..., X_k) \sim \mathcal{N}(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k, \sigma^2)$.

[8] An alternative could be to run quantile regressions and analyze each part of the distribution separately.

However, when the outcome also has many zero-valued observations (e.g., wealth), natural log transformations don't work well, as $\ln(0)$ is undefined. Instead, one can use the inverse hyperbolic sine (IHS or arcsinh) transformation:

$$\ln\left(Y + (Y^2+1)^{\frac{1}{2}}\right)$$

It approximates the natural logarithm (except for very small values of Y, it is $\approx \ln(2Y) = \ln(2) + \ln(Y)$), and so can be interpreted in the same way as a standard log-transformed dependent variable, while it is defined at zero, which allows us to retain zero-valued observations.

3.2 Inter/extrapolation assuming linear vs. exponential growth

How to interpolate/extrapolate a time series (e.g., in R) between years, assuming linear/exponential growth?

a. Choose whether linear/exponential growth
$\Longrightarrow$ This determines the growth formula (x_t as a function of x_0 and r) and hence the growth rate formula (r as a function of x_0 and x_t)

- Linear growth: $x_t = x_0 + r \times t \iff r = \frac{x_t - x_0}{t}$
- Exponential growth:
  - continuous: $x_t = x_0 e^{r_c t} \iff r_c = \frac{\ln(x_t) - \ln(x_0)}{t}$
  - discrete: $x_t = x_0(1+r_d)^t \iff r_d = \left(\frac{x_t}{x_0}\right)^{\frac{1}{t}}$

  Continuous and discrete exponential growth are approximately equivalent when the growth rate is small, i.e., $(1+r)^t \approx e^{rt}$ for small r.

b. Compute the growth rate r^{obs} using observations

- ⚠ *If exponential growth and our time series includes zeros:* $\ln()$ can't handle zero values, so we can't compute r... Two common workarounds are:
  - to add a small constant $\epsilon > 0$ before taking the logarithm, i.e., compute $\tilde{r}^{\text{obs}} = \frac{\ln(x_t+\epsilon) - \ln(x_0+\epsilon)}{t}$. However, this method is reasonable only if the zeros represent small or missing values rather than true zeros.
  - to use the inverse hyperbolic sine (IHS) function $\sinh^{-1}()$ as an approximation of $\ln()$, i.e., compute $\tilde{r}^{\text{obs}} \approx \frac{\sinh^{-1}(x_t) - \sinh^{-1}(x_0)}{t}$. Indeed, the IHS function behaves similarly to $\ln()$ for large values but remains well-defined for zero (see Figure 1).

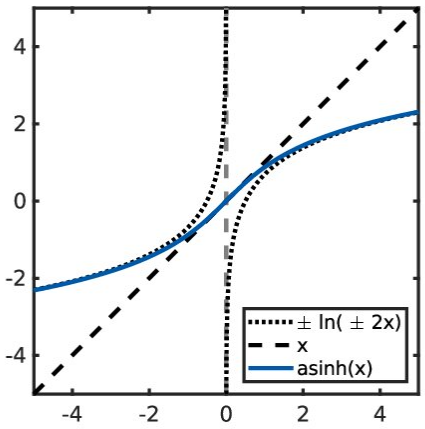


Figure 1: IHS approximation of the natural logarithm

c. Apply the growth formula embedding our r^{obs} to fill in missing observations.

4 Misc.

4.1 Kernel functions for non-parametric statistics

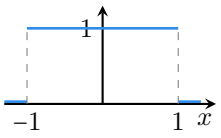
A **kernel** is a non-negative real-valued integrable function $k()$, used as a **weighting function** in non-parametric estimation techniques. They are also called "window functions" (notably in time-series).

Some applications require the function to satisfy additional conditions, for instance:

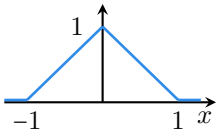
- normalization: $\int_{-\infty}^{+\infty} k(u)\,du = 1$
  In kernel density estimation, this ensures that the estimation produces a probability density function.
- symmetry: $\forall u,\ k(-u) = k(u)$
  This ensures that the average of the corresponding distribution is equal to that of the sample used.

Examples of commonly used symmetric kernels, with the arbitrary bounded support $[-1,1]$:

- uniform: $k(u) = \begin{cases} 1 & \text{if } |u| \le 1 \\ 0 & \text{otherwise} \end{cases}$



- triangular (Bartlett): $k(u) = \begin{cases} 1-|u| & \text{if } |u| \le 1 \\ 0 & \text{otherwise} \end{cases}$



- modified Bartlett (*a la Newey-West*): $k(u) = \begin{cases} 1 - \dfrac{u}{L+1} & \text{if } |u| \le L \\ 0 & \text{otherwise} \end{cases}$

- parabolic (Epanechnikov): $k(u) = \begin{cases} \dfrac{3}{4}\left(1-u^2\right) & \text{if } |u| \le 1 \\ 0 & \text{otherwise} \end{cases}$