Compact notes on probability and statistics

Erik Ø. Sørensen*

3rd edition, 2025

Contents

*FAIR, Department of Economics, NHH Norwegian School of Economics. erik.sorensen@nhh.no

1 Probability

Probability theory is defined in terms of set theoretic notions. We begin by defining a **sample space**, which is the set of all possible outcome of some event. Example, role of a die: $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$. We assign probability to **events**, which are subset of $\mathcal{S}$. The set of events that we can assign probabilities to is called a **σ-algebra**, which is a set of subsets of $\mathcal{S}$ that satisfies some properties, ensuring that probability measures are well-defined and consistent.

A σ-algebra $\mathcal{A}$, is a class of subsets of $\mathcal{S}$ with three requirements: We require that (i) if $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$; (ii) if $A \in \mathcal{A}$ and $B \in \mathcal{A}$ then $A \cup B$ is also in $\mathcal{A}$. This defines an **algebra**. We also require that (iii) if $A_1, A_2, \cdots \in \mathcal{A}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$, this last requirement is what restricts $\mathcal{A}$ to a σ-algebra.[1]

A function $P : \mathcal{A} \to \mathbb{R}$ is a **probability set function** if $P(A) \geq 0$ for all $A \in \mathcal{A}$; $P(\emptyset) = 0$; $P(\mathcal{S}) = 1$; and if $A_1, A_2, \ldots$ is a series of **disjoint** sets ($A_i \cap A_j = \emptyset$ when $i \neq j$), then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$. The triple $(\mathcal{S}, \mathcal{A}, P)$ is called a **probability space**.

With a discrete sample space, we will often work with the set of all subsets as the relevant σ-algebra. The set of all subsets is sometimes referred to as the **power set** and written $2^{\mathcal{S}}$. If the sample space is continuous (like the set of all real numbers),

[1] Some authors use the term *σ-field* to refer to a σ-algebra.

1. $P(A) = 1 - P(A^c)$.
2. $P(A) = P(A \cap B) + P(A \cap B^c)$.
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
4. $P(A \cap B) = P(A)P(B|A)$.
5. $P(A \cap B) \leq \min\{P(A), P(B)\}$.
6. $\mathrm{E}[\mathrm{E}[Y|X]] = \mathrm{E}[Y]$.
7. $\mathrm{E}[a + bX] = a + b\,\mathrm{E}[X]$.
8. $\mathrm{E}[aX + bY] = a\,\mathrm{E}[X] + b\,\mathrm{E}[Y]$.
9. $\mathrm{E}[g(X)] = \int g(x) f_X(x)\,\mathrm{d}x$ when the density exists.
10. $\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]$ when X and Y are independent.
11. $\mathrm{E}[|X + Y|] \leq \mathrm{E}[|X|] + \mathrm{E}[|Y|]$.
12. $\mathrm{var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$.
13. $\mathrm{var}[\overline{X}_n] = \mathrm{var}[X]/n$.
14. $\mathrm{cov}[X, Y] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$.
15. $\mathrm{cov}[X, X] = \mathrm{var}[X]$.
16. $\mathrm{var}[a + bX] = b^2\,\mathrm{var}[X]$.

Table 1: Standard normal distribution table

	s: combine with z to find cell entry $\Phi(z + s)$									
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.567	0.571	0.575
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.3	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.644	0.648	0.652
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.808	0.811	0.813
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839
1.0	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.858	0.860	0.862
1.1	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883
1.2	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.898	0.900	0.901
1.3	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.915	0.916	0.918
1.4	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.931	0.932
1.5	0.933	0.934	0.936	0.937	0.938	0.939	0.941	0.942	0.943	0.944
1.6	0.945	0.946	0.947	0.948	0.949	0.951	0.952	0.953	0.954	0.954
1.7	0.955	0.956	0.957	0.958	0.959	0.960	0.961	0.962	0.962	0.963
1.8	0.964	0.965	0.966	0.966	0.967	0.968	0.969	0.969	0.970	0.971
1.9	0.971	0.972	0.973	0.973	0.974	0.974	0.975	0.976	0.976	0.977
2.0	0.977	0.978	0.978	0.979	0.979	0.980	0.980	0.981	0.981	0.982
2.1	0.982	0.983	0.983	0.983	0.984	0.984	0.985	0.985	0.985	0.986
2.2	0.986	0.986	0.987	0.987	0.987	0.988	0.988	0.988	0.989	0.989
2.3	0.989	0.990	0.990	0.990	0.990	0.991	0.991	0.991	0.991	0.992
2.4	0.992	0.992	0.992	0.992	0.993	0.993	0.993	0.993	0.993	0.994
2.5	0.994	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995
2.6	0.995	0.995	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996
2.7	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
2.8	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
2.9	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.999	0.999	0.999

Note: To calculate $\Phi(x)$ for $x \geq 0$, find the row using the first decimal in x, and find the second decimal to find the correct column. For $x < 0$, use the symmetry of the distribution to find $\Phi(x) = 1 - \Phi(-x)$. The table is calculated with the R-function `pnorm` (R Core Team, 2023).

If (when) you find a typo, an error, or an inconsistency, please register an issue or a pull request at `https://github.com/ErikOSorensen/CompactNotes`.

This compact set of notes is designed to support the August Method Camp in probability and statistics for incoming PhD students at NHH Norwegian School of Economics. It is not intended to replace a comprehensive textbook or serve as a resource for independent study.

The notes draw heavily on two key references that I have used in teaching over the past decade: Hogg et al. (2013), a solid general textbook in mathematical statistics, and Linton (2017), which is more concise and specifically tailored for training econometricians. For the statistics component, econometrics textbooks will likely suffice for most students. However, for those seeking a deeper theoretical understanding of probability, a good starting point is Rosenthal (2006), while the rigor of Billingsley (1995) should meet the needs of any reader.

Wikipedia can also be a surprisingly helpful resource for topics in probability and statistics.

Although we will not cover software during the Method Camp, I recommend Wickham and Grolemund (2017) as a strong introduction to modern R, particularly for data wrangling in applied settings.

17. $\mathrm{var}[aX \pm bY] = a^2\,\mathrm{var}[X] + b^2\,\mathrm{var}[Y] \pm 2ab\,\mathrm{cov}[X,Y]$.
18. $\mathrm{cov}[X+Y,Z] = \mathrm{cov}[X,Z] + \mathrm{cov}[Y,Z]$.

References

Benjamin, Daniel J., et al. (2017). "Redefine statistical significance." *Nature Human Behaviour*, 2: 6–10.

Billingsley, Patrick (1995). *Probability and Measure*. Wiley, 3rd edition.

Haavelmo, Trygve (1944). "The probability approach in econometrics." *Econometrica*, 12: iii–115.

Hogg, Robert V., Joseph W. McKean, and Allen T. Craig (2013). *Introduction to Mathematical Statistics*. Pearson Education, 7th edition.

Leemis, Lawrence M. and Jacquelyn T. McQueston (2008). "Univariate distribution relationships." *American Statistician*, 62(1): 45–53.

Linton, Oliver (2017). *Probability, Statistics and Econometrics*. London, UK: Academic Press.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosenthal, Jeffrey S. (2006). *A First Look at Rigorous Probability Theory*. World Scientific, 2nd edition.

Wickham, Hadley and Garrett Grolemund (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly.

since the $\psi(n)/n$ term vanishes faster than x/n.

A useful formulation of **Taylor's formula** is that for continuous and at least twice differentiable functions functions $m\colon \mathbb{R} \to \mathbb{R}$, there exists a number ξ between 0 and t such that

$$m(t) = m(0) + m'(0)t + \frac{1}{2}m''(\xi)t^2.$$

The **indicator function** is defined as

$$1_{X \in S} = \begin{cases} 1 & \text{if } X \in S, \\ 0 & \text{if } X \notin S, \end{cases}$$

and is often useful for stating results (related to the "dummy variable" concept in applied work).

When $u = u(x)$ and $v = v(x)$, then the **product rule of differentiation** says that $(uv)' = u'v + uv'$, and **integration by parts** is the reverse statement—that

$$\int u \,\mathrm{d}v = uv - \int v \,\mathrm{d}u.$$

8 Some calculating rules for P, E, var and cov

Let A and B be some events; X, Y, and Z are random variables; a and b are constants.

there are technical difficulties working with the power set; the power set of $\mathbb{R}$ is not a σ-algebra. Instead, for real numbers we work with subsets that are open intervals, $\mathcal{B} = \{(a, b) \;:\; a, b \in \mathbb{R}\}$. Using all sets of type $\mathcal{B}$, and the countable unions and complements of these sets, we can define the **Borel σ-algebra**.

For A and B in $\mathcal{A}$, we define the **conditional probability** of A given B as

$$P(A|B) = P(A \cap B)/P(B),$$

when $P(B) > 0$. From the definition, we can find **Bayes rule** as

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

If A and B are events with positive probability, A and B are **independent** if and only if $P(A \cap B) = P(A) \cdot P(B)$, $P(A|B) = P(A)$, and $P(B|A) = P(B)$. These are equivalent conditions. Independence is symmetric but not transitive: If A and B are independent and B and C are independent, it does not follow that A and C are independent.

2 Random variables

A **random variable** (r.v.) is a function from a sample space to the real numbers, $X : \mathcal{S} \to \mathcal{S}_X$, with $\mathcal{S}_X \subset \mathbb{R}$. There are

mixed variables or variables we don't know the distribution of. To find a more general way to express expectations, we would need to extend our notion of *integration*, this is outside the scope of the method camp. But we sometimes see expressions such at $\mathrm{E}[X] = \int x \,\mathrm{d}F_X(x)$ or even $\mathrm{E}[X] = \int_\Omega X(\omega) P(\mathrm{d}\omega)$.

The k-th **moment** of X is $\mathrm{E}[X^k]$ (when it exists). We often write $\mu = \mathrm{E}[X]$ (the **mean**) and $\sigma^2 = \mathrm{E}[(X - \mu)^2]$ (the **variance**). The k-th **central moment** is $\mathrm{E}[(X - \mu)^k]$.

The expectation $m_X(t) = \mathrm{E}[\exp(tX)]$ is called the **moment generating function** (mgf) when it exists (is finite) in an open ball around zero. When the mgf exists, $m'(0) = \mathrm{E}[X]$, $m''(0) = \mathrm{E}[X^2]$ and $m^k(0) = \mathrm{E}[X^k]$, and the moment generating function completely characterizes the distribution of X. For a number of interesting distributions the moment generating functions does not exist (example: log-normal distribution). Even when the moment generating function does not exist, a variation called the **characteristic function** always exist, this is defined with complex numbers: $\phi_X(t) = \mathrm{E}[\exp(itX)]$, with $i = \sqrt{-1}$. The characteristic function completely characterizes any distribution.

2.3 Expectation results

1. Let X be a random variable and let m be a positive integer. Suppose that $\mathrm{E}[X^m]$ exists. If k is a positive integer and

Like for convergence in probability, we have some theorems about convergence in distribution:

1. If X_n converges to X in probability, then X_n converges to X in distribution.

2. If X_n converges to the constant b in distribution, then X_n converges to b in probability.

3. If X_n converges to X in distribution and Y_n converges in probability to 0, then $X_n + Y_n$ converges to X in distribution.

4. If X_n converges to X in distribution and g is a continuous function on the support of X, then $g(X_n)$ converges to $g(X)$ in distribution.

5. X_n, X, A_n, and B_n are random variables and a and b are constants. If X_n converges to X in distribution, $\mathrm{plim}(A_n) = a$, and $\mathrm{plim}(B_n) = b$, then

$$A_n + B_n X_n \xrightarrow{D} a + bX.$$

6.3 The Central Limit Theorem

There is a whole literature about generalizing central limit theorems. A basic one:

technical requirements: X must be **measurable**:

$$\mathcal{A}_X = \{A \subset \mathcal{S} \,:\, X(A) \in \mathcal{B}\} = \{X^{-1}(B) \,:\, B \in \mathcal{B}\} \subseteq \mathcal{A}.$$

Here $\mathcal{B}$ is a σ-algebra on $\mathcal{S}_X \subset \mathbb{R}$ and $\mathcal{A}$ is the σ-algebra on $\mathcal{S}$. Now the probability measure P_X is defined $P_X(B) = P(X^{-1}(B))$, and we have mapped the probability space $(\mathcal{S}, \mathcal{A}, P)$ to the probability space $(\mathcal{S}_X, \mathcal{B}, P_X)$.

To recognize the importance of measurability, consider the sample space $\mathcal{S}_e = \{a, b\}$, with the σ-algebra $\mathcal{A}_e = \{\emptyset, \mathcal{S}_e\}$, and the function $X_e : \mathcal{S}_e \to \mathbb{R}$, taking values $X_e(a) = 0$, $X_e(b) = 1$. This cannot be a random variable, since $X_e^{-1}(0) = \{a\}$ and $X_e^{-1}(1) = \{b\}$, and neither $\{a\}$ nor $\{b\}$ are in the σ-algebra $\mathcal{A}_e$. Measurability is a requirement on the combination of the function and the σ-algebra.

It is customary to use capital roman letters to refer to random variables and lower case letters to refer to particular values (numbers), such that $X(s) = x$ is the statement that at the point $s (\in \mathcal{S})$, the random variable X (a function) takes on the value x (a number, $x \in \mathbb{R}$).

2.1 Distribution

A random variable X has a **distribution function** (d.f.) F_X such that

$$F_X(s) = P_X((-\infty, s]) = P(\{a \,:\, X(a) \le s\}).$$

When the derivative of the distribution functionfor X exists, it is called the **density** (pdf) of X, written $f_X(s) = F'_X(s)$. We then say X is distributed **continuously**. General properties: $f_X(s) \ge 0$, $\int_a^b f_X(s)\,\mathrm{d}s = P(X \in [a, b])$, and $\int_{-\infty}^{\infty} f_X(s)\,\mathrm{d}s = 1$. When it is clear from the context, we drop the subscripts (and sometimes also the limits of integration): $\int f(s)\,\mathrm{d}s = 1$. If X and Y have distribution functions F_X and F_Y and $F_X(s) = F_Y(s)$ for all s, X and Y are **identically distributed**, but they are not necessarily the same random variables.

For **discrete** r.v., we define the **probability mass function** (pmf) as $p_X(s) = P(X \in \{s\})$ for $s \in \mathcal{S}$. This is analogous to the density function.

2.2 Expectation of a random variable

If X is a continuous random variable with pdf $f(x)$ and $\int_{-\infty}^{\infty} |x| f(x)\,\mathrm{d}x < \infty$, the **expectation** of X exists and is

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x.$$

If the set of outcomes is bounded ($\min(X) > -\infty$ and $\max(X) < \infty$), the expectation always exists. If Y is a discrete random variable with pmf p_Y, $\mathrm{E}[Y] = \sum_y y p_Y(y)$. Having different expressions for discrete and continuous random variables is awkward, since sometimes we need to develop theory for

Theorem 2 *Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution with mean μ and positive variance σ^2. Then the random variable*

$$Y_n = \left(\sum_{i=1}^{n} X_i - n\mu\right) / \sqrt{n}\sigma = \sqrt{n}(\overline{X}_n - \mu)/\sigma$$

converges in distribution to a random variable with a normal distribution $N(0, 1)$.

A simple proof can be constructed for the subset of cases where X has a moment generating function, the trick is to do a second order Taylor-approximation of the moment generating function of $X - \mu$.

The **Δ-rule** is often useful in conjunction with the CLT. It says that

Theorem 3 *If $\{X_n\}$ is a sequence of random variables such that $\sqrt{n}(\overline{X}_n - \theta)$ converges to $N(0, \sigma^2)$ in distribution, g is a differentiable function at θ, and $g'(\theta) \neq 0$, then*

$$\sqrt{n}\left(g(X_n) - g(\theta)\right) \xrightarrow{D} N\left(0, \left(g'(\theta)\right)^2 \sigma^2\right).$$

This can also be proven with a Taylor approximation.

The Δ-rule is the default approach to calculating the distribution of derived statistics. If we can show that the CLT applies

to some moments or parameters, and we are interested in a function of these moments or parameters, we can use the Δ-rule to calculate the distribution of these functions. Statistical packages will often do this automatically (example: Stata's `testnl` command).

7 Selected maths facts

In this section, there are some mathematical facts that turn out to be useful for probability and statistics but are not part of statistics itself.

The **exponential function**, written e^s or $\exp(s)$ can be defined by

$$e^x = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n,$$

or by the series expansion

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Remember that $\mathrm{d}e^s / \mathrm{d}s = e^s$.

If ψ is some function such that $\psi(n) \to 0$ as $n \to \infty$, then

$$\lim_{n\to\infty} \left(1 + \frac{x}{n} + \frac{\psi(n)}{n}\right)^n = e^x,$$

$k < m$, then $\mathrm{E}[X^k]$ exists.

2. Let $u(X)$ be a nonnegative function of the random variable X. If $\mathrm{E}[u(X)]$ exists, then for every positive constant c,
$$P(u(X) \geq c) \leq \mathrm{E}[u(X)]/c$$
(**Markov**).

3. Let the random variable X have a distribution with finite variance σ^2. Then for every $k > 0$,
$$P(|X - \mu| \geq k\sigma) \leq 1/k^2$$
(**Chebyshev**).

4. If ϕ is convex on an open interval Im and X is a random variable with finite expectation and support in I, then $\phi(\mathrm{E}[X]) \leq \mathrm{E}[\phi(X)]$ (**Jensen**).

2.4 Common distributions

There are some distributions that we should recognize:

Uniform A uniform distribution on $[a, b]$ is written $U(a, b)$, has a density $1/(b-a)$ on $[a, b]$ and a d.f. $F(x) = x/(b-a)$ on $[a, b]$. Expectation is $(a+b)/2$, the variance is $(b-a)^2/12$.

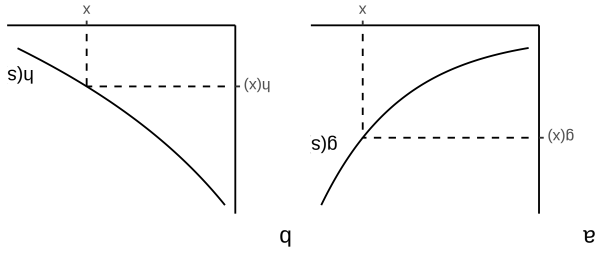

Figure 1: Functions of random variables

$f_Y(y) = f_X(g^{-1}(y)) \cdot |\,\mathrm{d}g^{-1}(y)/\,\mathrm{d}y|$. The absolute value of the derivative controls for whether g is increasing or decreasing.

For a function that is not one-to-one, but increasing on the subset A and decreasing on B, we can use the law of total probability,

$$F_Y(y) = P(Y \leq y) = P(Y \leq y, X \in A) + P(Y \leq y, X \in B),$$

and calculate each terms separately. See exercise 11 in the second set of exercises for an example of this.

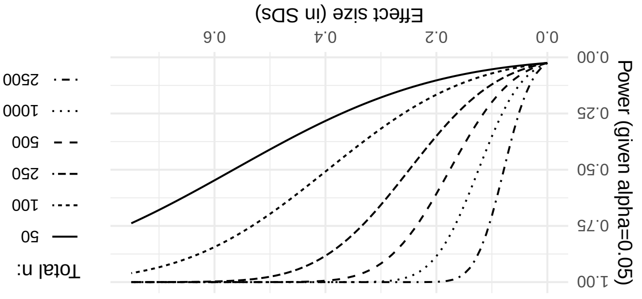

Figure 2: Power for a two-sided, two-sample t-test of $H_0 : \mu = 0$ as the effect size and sample size varies ($\alpha = 0.05$). Calculated with the R-command `power.t.test` (R Core Team, 2023).

In Figure 2 we see the power of a two-sided, two-sample t-test of $\mu_0 = 0$ at different alternative hypotheses about the effect size and at different total sample size. Measuring moderate effect sizes require large sample sizes to have the power that most consider sufficient (0.8 and above).

If we analyze classical hypothesis testing with Bayesian methods, we can uncover some unsettling consequences of how low power affects statistical inference. Consider Figure 3: we encode our prior knowledge as p, our belief that nature made the alternative hypothesis true, while $1 - p$ is our belief about the

1. Suppose $\operatorname{plim} X_n = X$ and $\operatorname{plim} Y_n = Y$. Then $\operatorname{plim}\left(X_n + Y_n\right) = X + Y$.

2. Suppose $\operatorname{plim} X_n = X$ and a is a constant. Then $\operatorname{plim} aX_n = aX$.

3. Suppose $\operatorname{plim} X_n = a$ and the function g is continuous at a. Then $\operatorname{plim} g(X_n) = g(a)$.

4. Suppose $\operatorname{plim} X_n = X$ and $\operatorname{plim} Y_n = Y$, then $\operatorname{plim} X_n Y_n = XY$.

The concept of probability limit is closely tied to the statistical concept of consistency. Let X be a r.v. with d.f. $F(s, \theta)$, for some $\theta \in \Omega$. Let $X_1, X_2, \ldots, X_n$ be a random sample on X, and let T_n be a statistic. T_n is a **consistent** estimator of θ if $\operatorname{plim} T_n = \theta$.

6.2 Convergence in distribution

$\{X_n\}$ is a sequence of random variables and X is a random variable, let F_{X_n} and F_X be the respective distribution functions. Let $C(F_X)$ be the set of points at which F_X is continuous. Now X_n **converges in distribution** to X if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for all $x \in C(F_X)$, and we sometimes write

$$X_n \xrightarrow{D} X.$$

Standard normal A standard normal distribution is written $N(0, 1)$, the density is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

The distribution function is often written Φ but there is no analytical expression for this. See Table 1 in the back for some tabulated values. The moment generating function is $m(t) = \exp(t^2/2)$. The expectation is 0 and the variance is 1. The standard normal distribution is **symmetric** around zero, $\Phi(s) = 1 - \Phi(-s)$ for all $s \in \mathbb{R}$.

Exponential An exponential distribution with parameter λ is written $\mathrm{Exp}(\lambda)$, has density $f(x) = \lambda \exp(-\lambda x)$ on $[0, \infty)$, and distribution function $F(x) = 1 - \exp(-\lambda x)$. The expectation is $1/\lambda$ and the variance is $1/\lambda^2$.

Binomial If an event happens with probability p in a single (binary) trial, the distribution of the number of events k in n trials has probability mass function $p(k) = \binom{n}{k} p^k (1-p)^{n-k}$ (with the binomial coefficient $\binom{n}{k} = n!/k!(n-k)!$). The expectation is np and the variance is $np(1-p)$. The special case $n = 1$ is often very useful (**Bernoulli**).

Poisson A Poisson distribution on the integers $\{0, 1, 2, \dots\}$ has probability mass function $p(x) = \lambda^x \exp(-\lambda)/x!$, the

distribution function is not pretty. The moment generating function is $m(t) = \exp(\lambda(\exp(t) - 1))$. The expectation is λ and the variance is λ.

A more complete overview of distributions is given by Leemis and McQueston (2008). There are also multi-volume handbooks written with properties of various distributions.

2.5 Functions of random variables

If X is a random variable and g is a function, $Y = g(X)$ is another random variable.

In Figure 1a, $Y = g(X)$ is an increasing function (one-to-one). Considering the distribution of Y, the distribution function for Y is $F_Y(y) = P(Y \leq y)$. We can substitute the definition of Y into this: $F_Y(y) = P(g(X) \leq y)$. We can now apply the inverse of g on both sides of of the inequality in P, relying on g being increasing: $P(g(X) \leq y) = P(g^{-1}(g(X)) \leq g^{-1}(y)) = P(X \leq g^{-1}(y))$, and we can conclude that $F_Y(y) = F_X(g^{-1}(y))$. The density follows by differentiation, $f_Y(y) = f_X(g^{-1}(y))\, \mathrm{d}g^{-1}(y)/\,\mathrm{d}y$.

If instead $Y = h(X)$ and h is decreasing and one-to-one (as in Figure 1b), we start with a different inequality: If $h(X) \leq y$, then $X \geq h^{-1}(y)$, and $F_Y(y) = P(X \geq h^{-1}(y)) = 1 - P(X \leq h^{-1}(y)) = 1 - F_X(h^{-1}(y))$. Either way, for increasing or decreasing transformations g, taking derivatives we find that

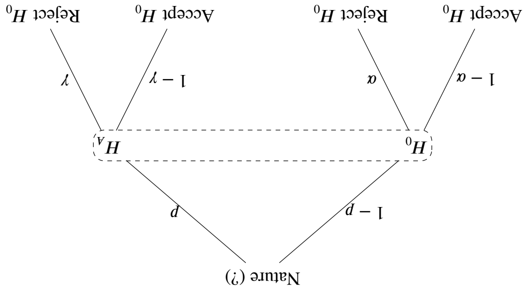

Figure 3: Statistical game tree

probability that H_0 is true. We cannot tell if we are at the H_0 or at the H_A node, but if we are at H_0 we reject with probability α, the level of significance for our test (a type-I error). If we are at H_A, we reject with probability γ, the power of the test. Rejecting H_0, the Bayesian posterior belief in H_A is

$$P(H_A|\text{Reject } H_0) = \frac{P(\text{Reject } H_0|H_A) \cdot P(H_A)}{P(\text{Reject } H_0)}, \\ = \frac{p\gamma}{(1-p)\alpha + p\gamma}.$$

Consider a surprising ($p = 0.1$) rejection of H_0 in an under-powered study ($\gamma = 0.5$), our posterior belief in H_A should not

be very strong: $0.1 \cdot 0.5/(0.9 \cdot 0.05 + 0.1 \cdot 0.5) = 0.53$. For this reason some people argue for a much stricter α: Benjamin et al. (2017) argue that we should use $\alpha = 0.005$ as a conventional level instead of 0.05 (in this example, if we had the surprising result at $\alpha = 0.005$, that would give us $P(H_A|\text{Reject } H_0) = 0.92$).

6 Asymptotic theory

6.1 Convergence in probability

$\{X_n\}$ is a sequence of random variables, and X is a random variable, both defined on the same sample space. Now X_n **converges in probability** to X if, for all $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0.$$

We say that $\operatorname{plim} X_n = X$, or that $X_n \xrightarrow{P} X$.

Theorem 1 *Let* $\{X_n\}$ *be sequence of iid random variables with common mean* μ *and finite variance* σ^2. *If* $\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$, *then* $\operatorname{plim} \overline{X}_n = \mu$. *(Weak law of large numbers.)*

There are a number of results about how probability limits work:

if $\mathrm{E}[X] = \mu_0$ using an expression such as

$$t_n = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}.$$

If we know σ, then t_n would be normally distributed $N(0, 1)$ under H_0 : $\mathrm{E}[X] = \mu_0$. This would be true even in small samples. We can modify this most simple t-test in various ways.

When we control the type-I error probability to α and the type-II error probability to β, we say that the **power** is $1 - \beta$. In order to calculate power, it is necessary to be explicit about the alternative hypothesis we consider—or calculate for a range of alternative hypotheses. Power is the probability of a significant result under the alternative hypothesis, and is a good thing. Among people who design experiments, it is customary to aim for a power of 80% or 90%.

Assume a simple one-sample, one-sided t-test against a null-hypothesis of $\mathrm{E}[X] = \mu_0 = 0$. This will be significant at $\alpha = 5\%$ when $t > 1.64$, with probability

$$P\left(\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} > 1.64 | E[X] = \mu_A\right) = 1 - \Phi\left(1.64 - \frac{\mu_A}{\sigma/\sqrt{n}}\right).$$

Only the ratio $\delta = (\mu_A - \mu_0)/\sigma$, known as the **effect size**, and the sample size n matters.

3 Random vectors

Given a sample space $\mathcal{S}$. Consider the two random variables X_1 and X_2, which assign to each element s of $\mathcal{S}$ an ordered pair of numbers $X_1(s) = x_1$, $X_2(s) = x_2$. Then we say that (X_1, X_2) is a **random vector**.

If (X_1, X_2) has density $f_{X_1,X_2}(x_1, x_2)$, the **marginal density** of X_1 is $f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)\,\mathrm{d}x_2$.

If (X_1, X_2) is discrete, then the conditional probability of X_1 given X_2 is

$$p_{X_1|X_2}(x_1|x_2) = \frac{p_{X_1,X_2}(x_1, x_2)}{p_{X_2}(x_2)}, \quad \text{where } p_{X_2}(x_2) > 0.$$

But if (X_1, X_2) is continuous, all point probabilities are zero, and another way to determine conditional distribution is necessary. It turns out that if (X_1, X_2) has a density, it is useful to define the **conditional density** of X_1 given X_2 as

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}, \quad \text{where } f_{X_2}(x_2) > 0.$$

We can now define the **conditional expectation**, $\mathrm{E}[X_2|X_1 = x_1] = \int_{-\infty}^{\infty} s f_{X_2|X_1}(s|x_1)\,\mathrm{d}s$.

For the completely general case (which we won't be working

sample.

$$\mathrm{E}[g(X;\theta)] = 0, \qquad \text{(theory)}$$

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i;\widehat{\theta}) = 0. \qquad \text{(analog)}$$

The second expression is a concrete equation that can be solved by analytical or computational methods for $\widehat{\theta}$ as a function of the data $(x_1, x_2, \dots, x_n)$. An example: To estimate the expectation of X, choose $g(X;\mu) = X - \mu$. The solution is $\widehat{\mu} = n^{-1}\sum X_i$, often written $\widehat{\mu} = \overline{X}_n$. $\widehat{A}$ always mean the **estimate** of A, and is a number (or a vector of numbers). The *function* to calculate the estimate based on a random sample (data) is called the **estimator**. The parameter we aim to estimate is called the **estimand**.

Similarly, for a simple regression model,

$$Y_i = \alpha + \beta X_i + \varepsilon_i,$$

the core assumption is often that $\mathrm{cov}[\varepsilon_i, X_i] = 0$. Reformulating, $\mathrm{cov}[Y_i - (\alpha + \beta X_i), X_i] = 0$, or $\mathrm{cov}[Y_i, X_i] = \beta\,\mathrm{var}[X_i]$. The empirical analog of this is

$$\widehat{\mathrm{cov}[X_i, Y_i]} = \widehat{\beta}\widehat{\mathrm{var}[X_i]}.$$

This equation is easily solved for $\widehat{\beta} = \widehat{\mathrm{cov}[X_i, Y_i]}/\widehat{\mathrm{var}[X_i]}$. To find $\widehat{\alpha}$ use the same kind of procedure for the theoretical moment

$\mathrm{E}[\varepsilon_i] = 0$. This kind of application of the analog principle is called **method of moments**. Note that in this regression case, the assumption $\mathrm{cov}[Y_i - (\alpha + \beta X_i), X_i] = 0$ is equivalent to the expected first order condition of a least-square minimization.

Often what seems like different approaches end up as equivalent to method of moment estimation. An example: With **maximum likelihood**, under some regularity assumptions maximum likelihood estimation is equivalent to method of moments with the additional result that the expectation of the **score**, the first order condition of the log likelihood, is zero. But in general, maximum likelihood is based on maximizing (with respect to the parameters θ) the probability (or density) of the data, $L(\boldsymbol{X};\theta)$, given the data $\boldsymbol{X}$. Most of the time, we prefer to maximize the log likelihood, which for an iid sample is just the sum of individual log likelihoods.

In most of applied economics, we think of the parameters we want to estimate as properties of some **data generating process** that potentially can generate infinitely many realizations and which is inherently stochastic. This tradition goes back to Trygve Haavelmo (1944). In survey statistics we instead formulate the problem such that there is a finite population true and deterministic value, but our estimates are random variables because we only sample a subset of the full population. Either way, our estimates are random variables, and the study of the stochastic properties of these random variables is known as with), we require that for all Borel subsets $S \subset \mathbb{R}$,

$$\mathrm{E}[P(A|X) \cdot 1_{X \in S}] = P(A \cap \{X \in S\}),$$
$$\mathrm{E}[\mathrm{E}[Y|X] \cdot 1_{X \in S}] = \mathrm{E}[Y \cdot 1_{X \in S}].$$

These equations tie down both conditional probability and conditional expectation. Note that with the indicator function, it would be possible to *define* probability as $P(X \in A) = \mathrm{E}[1_{X \in A}]$, this would push the problem of defining conditional probability into the theory of integration.

Let (X_1, X_2) have a joint distribution function $F(x_1, x_2)$ and let X_1 and X_2 have marginal distribution functions $F_1(x_1)$ and $F_2(x_2)$. Then X_1 and X_2 are **independent** if and only if

$$F(x_1, x_2) = F_1(x_1)F_2(x_2),$$

the joint distribution function being the product of marginal distribution functions. It follows that if the corresponding densities exist, X_1 and X_2 are independent if the joint density is the product of marginal densities.

If X_1 and X_2 are independent and $\mathrm{E}[u(X_1)]$ and $\mathrm{E}[v(X_2)]$ exist, then it is the case that $\mathrm{E}[u(X_1)v(X_2)] = \mathrm{E}[u(X_1)] \cdot \mathrm{E}[v(X_2)]$.

If X and Y are random variables, $\mathrm{E}[\mathrm{E}[Y|X]] = \mathrm{E}[Y]$, known as the law of **iterated expectations**.

For complete generality, let $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$ be one-to-one transformed random variables, and

let w_1 and w_2 be inverses. Then in general,

$$\iint_A f_{X_1,X_2}(x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2$$
$$= \iint_B f_{X_1,X_2}(w_1(y_1, y_2), w_2(y_1, y_2))|\det(\mathbf{J}(y_1, y_2))|\,\mathrm{d}y_1\,\mathrm{d}y_2,$$

where $\mathbf{J}(y_1, y_2)$ is the Jacobian of $\mathbf{w}(\mathbf{y})$ and B is the image $B = \mathbf{u}(A)$. This is an application of the general formula for integration with change of variables in calculus.

A vector $\mathbf{X} = (X_1, \dots, X_n)$ where each element is an *independent* draw from the same distribution as the random variable X is known as a **random sample** of size n of the variable X.

4 Estimation

For estimation, we would like to have some rules to uncover empirical analogs to theoretically defined parameters. We will assume we have access to a random sample from some defined population (with dependency in data, we often start by a transforming data such that the transformed data is a random sample). The **analog principle** is to start with a theoretically defined parameter and set its empirical analog equal to the same in the

inference, which is a crucial input into hypothesis testing.

Often we want our estimators to be **unbiased**, such that $\mathrm{E}[\hat{\theta}] = \theta_0$, with θ_0 being the true value. We can define the **bias** as $b = \mathrm{E}[\hat{\theta}] - \theta_0$. Even if estimators are biased (most non-linear estimators are biased), they can be **consistent**. To define consistency we need some asymptotic theory (in Section 6).

5 Hypothesis testing

Assume that we have two alternative ways to think of what might be true.

The distribution of X is $f(x;\theta)$, with parameter $\theta \in \Omega$, and our ideas about truth can be described by regions of Ω:

$$H_0 : \theta \in \omega_0 \quad \text{vs } H_A : \theta \in \omega_A.$$

Our task is, based on a sample from the distribution of X, determine whether to keep insisting that H_0 is true or reject that in favor of H_A. What we want is often to fix the probability of Type I error ("significance level") to $\alpha = 0.01, 0.05, \dots$, and conditional on that, minimize probability of type II error. By **Type I error** we mean that a null hypothesis is rejected when the null is true; by **Type II error** we mean that the null hypothesis is not rejected even though it is false.

Strictly speaking, we decide on a regions of the outcome space, $(X_1, \dots, X_n) \in C$, in which to accept or reject H_0, but in practice we define test-statistics $T_n(X_1, \dots, X_n)$ and decide on **critical regions** of the test statistics. Instead of reporting reject/accept with a given α-criterion (often indicated with stars in tables), papers often report p-values.

P-values are the probability of at least as extreme data given H_0,

$$P(\text{data}|H_0).$$

P-values do not address the likelihood of H_0 being true. In classical inference, probabilities are *never* attached to hypotheses: Hypotheses are true or false, and this is not a sampling issue. If we want to be Bayesian about hypotheses (have beliefs about them), we need to incorporate the priors:

$$P(H_0|\text{data}) = \frac{P(\text{data}|H_0)P(H_0)}{P(\text{data})}.$$

Very little of applied economics is explicitly Bayesian, most is anchored in **frequentist inference**. In frequentist philosophy, there is always a true (but unknown) value of any parameter, and we do not put probabilities on our beliefs about this parameter. Probabilities are restricted to the **sampling distribution** of our estimators *under some null hypothesis*.

If we assume that X is normally distributed, we could test