2 Statistical Inference

As mentioned in the last chapter, in econometrics, we interpret data y as draws of a random variable Y from probability distribution $p(Y|\theta)$ indexed by a parameter θ. This distribution is usually specified based on economic theory or a hypothesized relationship among different variables included in y. In some cases, we only specify certain aspects of this distribution, e.g. the (conditional) mean.

Suppose we observe a particular realization y. What can we say about θ? This is the problem of statistical inference, discussed in this chapter. The discussion here covers the statistical inference problem both from a frequentist/classical – in Section 2.1 – as well as a Bayesian point of view – in Section 2.2 –, each time discussing point estimation, including the evaluation of point estimators in both finite and asymptotic samples, hypothesis testing and the construction of coverage sets, called confidence sets in the frequentist paradigm and credible sets in the Bayesian paradigm.

To convey the fundamentals in a simple setting, unless otherwise stated, this chapter assumes that θ is a scalar. Furthermore, we assume that we have a data sample $\{y_i\}_{i=1}^n$ of n observations (realizations) y_i of the same underlying RV Y_i which are drawn independently from some distribution $p(y|\theta)$ (i.e. the observations are i.i.d.). Thereby, Y_i (and hence y_i) is assumed to be a scalar. We take our observations together into the $n \times 1$ vector $y = (y_1, ..., y_n)'$.

As a simple running example, let y_i denote the height (in centimeters) of individual i, and suppose we have data on the height of n adult females in Switzerland. We are interested in estimating the average height of all adult females in the population (i.e. in all of Switzerland), denoted by θ. We can write

$$y_i = \theta + u_i \ , \tag{2.1}$$

where u_i is a mean-zero error: $\mathbb{E}[u_i|\theta] = 0$. Note that this implies $\mathbb{E}[y_i|\theta] = \theta$.[1] For now, this is the only assumption we make on the pdf $p(y|\theta)$ that generated our data y given the parameter θ.

The previous chapter explicitly distinguished the cdf $F_Y(y)$, the pdf for continuous RVs $f_Y(y)$ and the probability function for discrete RVs $\mathbb{P}[Y = y]$ or $\mathbb{P}[y]$. In this and subsequent chapters we will predominantly deal with continuous random variables and write $p(y)$ for the pdf (and $F(y)$ or $\mathbb{P}[Y \leq y]$ for the cdf, if needed).

2.1 Frequentist/Classical Inference

2.1.1 Point Estimation

Definition 22. *A point estimator $\delta(Y)$ is a mapping from the sample space of Y to the parameter space Θ: $\delta : \mathcal{Y} \to \Theta$.*

In other words, for any realization of Y, $y \in \mathcal{Y}$, the point estimator gives us a value $\delta(y) \in \Theta$. Loosely speaking, $\delta(y)$ is our "best guess" where θ lies given that we observed $Y = y$. Typically, we denote the point estimator as $\hat{\theta}$, but it is crucial to understand that the point estimator of a parameter θ is a function of the data Y, hence the definition uses $\delta(Y)$, where δ stands for "decision". While we think of the point estimator $\delta(Y)$ as a function of the RV Y, if we evaluate it at the particular realization y that we observe, we call the result an estimate rather than an estimator.[2]

There are different estimators that exploit different aspects of $p(y|\theta)$, the probability distribution where the data $y = (y_1, ..., y_n)'$ (supposedly) came from. Our discussion starts with the classical case where θ is treated as a fixed but unknown parameter. For this reason, we condition the moments and distributions on θ.

The Least Squares (LS) approach constructs a point estimator by minimizing the sum of squared deviations of y_i from their (conditional) mean $\mathbb{E}[Y_i|\theta]$:

$$\hat{\theta}_{LS} \equiv \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} (y_i - \mathbb{E}[Y_i|\theta])^2 \ .$$

In our simple case, $\mathbb{E}[Y_i|\theta] = \theta$, which leads to $\hat{\theta}_{LS} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Note that the LS estimator

[1] Note that this assumption specifies the mean of y, not only that of an individual observation y_i.

[2] In other words, the point estimator is a rule how we form our best guess based on a RV Y, whereas the point estimate is the best guess we obtain given a particular observation y of Y.

only requires assuming a functional form for the expectation $\mathbb{E}[Y_i|\theta]$, i.e. we do not need to know the full distribution of $Y_i|\ \theta$.

The Method of Moments (MM) approach constructs a point estimator by setting empirical moments of $\{y_i\}_{i=1:n}$ equal to the corresponding population moments in $p(Y_i|\theta)$. For example, using the mean, we get the same result as under LS, i.e.

$$\mathbb{E}[Y_i|\hat{\theta}_{MM}] = \frac{1}{n}\sum_{i=1}^{n} y_i \ .$$

Because, in our simple case, $\mathbb{E}[Y_i|\theta] = \theta$, we obtain once again $\hat{\theta}_{MM} = \frac{1}{n}\sum_{i=1}^{n} y_i$. It is important to realize that the MM estimator can be constructed also based on higher-order moments, which, however, typically involves more assumptions. In our example, assuming $\mathbb{V}[Y|\theta] = 1$, we obtain the second moment of the distribution of $Y_i|\theta$: $\mathbb{E}[Y_i^2|\hat{\theta}] = \mathbb{V}[Y_i|\theta] + \mathbb{E}[Y_i|\hat{\theta}] = 1 + \theta^2$. In turn, we can define the MM estimator $\hat{\theta}_{MM}$ also as the quantity that equalizes this second moment in our sample and in the population:

$$\mathbb{E}[Y_i^2|\hat{\theta}_{MM}] = \frac{1}{n}\sum_{i=1}^{n} y_i^2 \quad \Rightarrow \quad \hat{\theta}_{MM} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - 1} \ .$$

Similarly to LS, the MM estimator does not require the researcher to specify the whole distribution of $Y_i|\theta$, but only a couple of moments.[3]

The Maximum Likelihood (ML) approach constructs a point estimator by maximizing the likelihood (i.e. the probability) of observing our particular sample $y = (y_1, ..., y_n)'$ out of all the possible draws of $Y = (Y_1, ..., Y_n)'$:

$$\hat{\theta}_{ML} \equiv \arg\max_{\theta\in\Theta} \mathcal{L}(\theta|y) = \arg\max_{\theta\in\Theta} \ell(\theta|y) \ ,$$

where the likelihood $\mathcal{L}(\theta|y) = p(y|\theta)$ is defined to be the pdf of our observed data y given the parameter θ, and $\ell(\theta|Y) = log\ \mathcal{L}(\theta|Y)$ is its natural logarithm. Because our observations y_i are independent, we can write

$$\mathcal{L}(\theta|y) = p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta) \ .$$

In order to proceed with ML estimation, we need to assume the whole distribution $p(y_i|\theta)$

[3] We need to specify enough moments so that we can solve for θ. If θ is a scalar, as here, then one (scalar) moment suffices. If θ were two dimensional (and Y_i still a scalar), then we would need to specify two (scalar) moments, e.g. $\mathbb{E}[Y_i|\theta]$ and $\mathbb{V}[Y_i|\theta]$.

where our observations come from. For example, we might say that the error term in Eq. (2.1) is Normally distributed with a variance of one: $u_i \overset{i.i.d.}{\sim} N(0,1)$, which in turn implies $Y_i|\theta \overset{i.i.d.}{\sim} N(\theta,1)$ and yields the likelihood

$$\mathcal{L}(\theta|y) = \prod_{i=1}^{n} (2\pi)^{-\frac{1}{2}} exp\left\{\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^2\right\} .$$

In turn, this leads to the log-likelihood

$$\ell(\theta|y) = \frac{n}{2}\ln(2\pi) + \frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^2 ,$$

which, once again, gives $\hat{\theta}_{ML} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

The notation $\mathcal{L}(\theta|y)$ shows that under ML estimation, the pdf $p(y|\theta)$, a function of data/realizations y given the parameter θ, is interpreted as a function of θ given y, which allows us to find an optimal θ – the one that maximizes the likelihood – given the observed data y. A useful property of the ML estimator is invariance: if $\hat{\theta}$ is the ML estimator of θ, then $f(\hat{\theta})$ is the ML estimator of $f(\theta)$.

Finite Sample Properties We can analyze the properties of a point estimator $\hat{\theta}$ for a given sample size n. Recall that $\hat{\theta}$ is a function the data Y, e.g. we obtained $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ above. As a result, we can compute the moments of $\hat{\theta}$ based on the moments of Y_i, while knowing the whole distribution of Y_i allows us to find the distribution of $\hat{\theta}$.

Definition 23. *A point estimator $\delta(Y)$ of θ is unbiased if $\mathbb{E}[\delta(Y)|\theta] = \theta$.*

The expectation of the point estimator, $\mathbb{E}[\hat{\theta}|\theta]$, tells us the average value we expect to get if we were to randomly draw different samples $y = (y_1, ..., y_n)'$ of size n and compute the point estimator $\delta(y)$ for each of them. To calculate $\mathbb{E}[\delta(Y)|\theta]$, we only need to specify the expectation of our data $\mathbb{E}[Y_i|\theta]$; no other moment of $p(Y_i|\theta)$ is needed. In our running example, $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is unbiased:

$$\mathbb{E}[\hat{\theta}|\theta] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i|\theta] = \frac{1}{n}\sum_{i=1}^{n}\theta = \theta .$$

In contrast, the estimator $\hat{\theta}_* = \frac{1}{n-1}\sum_{i=1}^{n} Y_i$ is not unbiased:

$$\mathbb{E}[\hat{\theta}_*|\theta] = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n} Y_i\right] = \mathbb{E}\left[\frac{n}{n-1}\hat{\theta}\right] = \frac{n}{n-1}\theta \neq \theta .$$

Provided that we specify also the variance of our observations, $\mathbb{V}[Y_i|\theta]$, we can also compute the variance of our point estimator, $\mathbb{V}[\delta(Y)|\theta]$. The quantity $\mathbb{V}[\delta(Y)|\theta]$ tells us the dispersion we would get if we were to randomly draw different samples $y = (y_1, ..., y_n)'$ of size n and compute the point estimator $\delta(y)$ for each of them. In our running example assuming $\mathbb{V}[Y_i|\theta] = 1$, we get

$$\mathbb{V}[\hat{\theta}|\theta] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[Y_i|\theta] = \frac{1}{n^2}\sum_{i=1}^{n} 1 = \frac{1}{n}$$

and

$$\mathbb{V}[\hat{\theta}_*|\theta] = \mathbb{V}\left[\frac{n}{n-1}\hat{\theta}\right] = \frac{n^2}{(n-1)^2}\mathbb{V}\left[\hat{\theta}\right] = \frac{n}{(n-1)^2} \ .$$

Definition 24. *A point estimator $\delta_1(Y)$ of θ is more efficient than the point estimator $\delta_2(Y)$ if $\mathbb{V}[\delta_1(Y)|\theta] < \mathbb{V}[\delta_2(Y)|\theta]$.*

In our example, $\hat{\theta}$ is more efficient than $\hat{\theta}_*$ because $n^{-1} < n/(n-1)^2$ for any $n \geq 1$.

Going another (big) step further, if we are willing to make an assumption not only on the mean adn variance of our data $Y_i|\theta$, but on the whole distribution $p(Y_i|\theta)$, we might be able to deduce not only the mean and the variance of the point estimator $\delta(Y)$, but its whole distribution. In our example, if $Y_i|\theta \overset{i.i.d.}{\sim} N(\theta, 1)$, then we know that $\hat{\theta}$ is itself Normally distributed, because it is the average of Normally distributed RVs (see Appendix B):

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N\left(\theta, n^{-1}\right) \ .$$

The same holds for $\hat{\theta}_*$, which follows a Normal distribution with the above moments $\mathbb{E}[\hat{\theta}_*|\theta]$ and $\mathbb{V}[\hat{\theta}_*|\theta]$.

Asymptotic Properties We can also analyze the properties of a point estimator $\hat{\theta}$ as the sample size $n \to \infty$. This gives us a sense of how our estimator behaves under large samples.

Definition 25. *A point estimator $\delta(Y)$ of θ is consistent if $\delta(Y) \xrightarrow{p} \theta$.*

Recall the definition of convergence in probability, applied to this case: $\forall\, \epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}[|\delta(Y) - \theta| > \epsilon] = 0$, i.e. the probability of obtaining a point estimate $\delta(y)$ that is further away from θ than the distance ϵ converges to zero, no matter how small we make ϵ. In other words, consistency tells us that our point estimator "gets it right" in the limit, for a large sample. While we sometimes do not have – or do not want (see below) – an unbiased estimator, we typically require an estimator to be consistent.

We know that $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n Y_i$ is consistent, because $\frac{1}{n}\sum_{i=1}^n Y_i \xrightarrow{p} \mathbb{E}[Y_i|\theta] = \theta$ by the WLLN. The estimator $\hat{\theta}_* = \frac{1}{n-1}\sum_{i=1}^n Y_i$ is also consistent:

$$\hat{\theta}_* = \frac{1}{n-1}\sum_{i=1}^n Y_i = \frac{n}{n-1}\hat{\theta} \xrightarrow{p} \theta \ ,$$

because $\hat{\theta} \xrightarrow{p} \theta$ and $\lim_{n\to\infty} \frac{n}{n-1} = 1$, and the Slutsky theorem tells us that we can form the limit of their product as the product of the limits.

Based on the CLT, we also know the asymptotic distribution of $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n Y_i$:

$$\sqrt{n}\left(\hat{\theta} - \theta\right)|\theta \xrightarrow{d} N(0,1) \ ,$$

because $\mathbb{V}[Y_i|\theta] = 1$. While for the finite sample distribution we had to know the distribution of the data $Y_i|\ \theta$, we found this asymptotic distribution under minimal assumptions: independence and identical distribution of our observations $\{y_i\}_{i=1}^n$. Therefore, if we do not know the distribution of the data, we can rely on the CLT to obtain the asymptotic distribution of our estimator. In turn, we can approximate the finite sample distribution based on the asymptotic one, by saying that

$$\sqrt{n}\left(\hat{\theta} - \theta\right)|\theta \overset{approx.}{\sim} N(0,1) \quad \text{or, equivalently,} \quad \hat{\theta}|\theta \overset{approx.}{\sim} N(\theta, n^{-1}) \ ,$$

already for our finite sample of size n.[4]

Evaluation of Estimators In finite samples, point estimators δ can be evaluated and compared to other point estimators using a loss function $L(\theta,\delta)$. Common choices are the quadratic/L2 loss function $L(\theta,\delta) = (\theta-\delta)^2$ or L1 loss function $L(\theta,\delta) = |\theta - \delta|$.

Definition 26. *Frequentist risk:* $R(\theta,\delta) = \mathbb{E}[L(\theta,\delta(Y))|\theta] = \int_y L(\theta,\delta(Y))p(y|\theta)dy$.

Frequentist risk takes the expectation of the loss function treating Y as a RV and θ as fixed. This means that it determines the behavior of an estimator conditional on a true θ and under the assumption that nature provides us repeatedly with draws from Y. For example,

[4] In our particular example with $Y_i|\theta \sim N(\theta,1)$, by approximating the finite sample distribution via the asymptotic one, we get it exactly right. However, this does not need to be the case. For example, let $Y_i|\ \theta \sim Exp(\lambda)$, with $\mathbb{E}[Y_i|\lambda] = \lambda^{-1}$ and $\mathbb{V}[Y_i|\lambda] = \lambda^{-2}$, and suppose that – for some reason – we form the estimator $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^n Y_i$. Then $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^n Y_i \sim Gamma(n, n\lambda)$, while $\sqrt{n}\lambda\left(\hat{\lambda} - \lambda^{-1}\right)|\lambda \xrightarrow{d} N(0,1)$, i.e. $\hat{\lambda}|\lambda \overset{approx.}{\sim} N(\lambda^{-1}, \lambda^{-2}n^{-1})$.

if we want to estimate θ based on a single observation of the RV $Y|\theta \sim N(\theta, 1)$ and we use a quadratic loss function $L = (\theta - \delta)^2$, we get the following frequentist risk for a point estimator of the form $\delta = cY$:

$$R(\theta, \delta) = \mathbb{E}[(\theta - cY)^2|\theta] = \theta^2 - 2c\theta\mathbb{E}[Y|\theta] + c^2\mathbb{E}[Y^2|\theta] = \theta^2(1-c)^2 + c^2 \; .$$

In this case, it is clear that $c = 1$ and therefore the point estimator $\delta(Y) = Y$ minimize frequentist risk.[5] In general, however, R depends on θ, which is unknown, which means that there is oftentimes no unique ranking of estimators independent of the supposedly true value of θ. Possible rankings are based on the minimax risk or inadmissibility. These concepts are discussed in the Appendix.

Under the quadratic loss function, we can decompose frequentist risk into a bias- and a variance-term. If θ is a scalar, we have $L(\theta, \delta) = (\theta - \delta)^2$ and

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}[(\theta - \delta)^2|\theta] \\ &= \theta^2 - 2\theta\mathbb{E}[\delta|\theta] + \mathbb{E}[\delta^2|\theta] \\ &= (\theta - \mathbb{E}[\delta[\theta])^2 + \mathbb{E}[\delta^2[\theta] - \mathbb{E}[\delta|\theta]^2 \\ &= (\theta - \mathbb{E}[\delta[\theta])^2 + \mathbb{V}[\delta|\theta] \; , \end{aligned}$$

i.e. $R = \text{bias}^2 + \text{variance}$. If θ is a k-dimensional vector, let W be any positive-defiite weighting matrix and define $L(\theta, \delta) = (\theta - \delta)'W(\theta - \delta)$. We get

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}[(\theta - \delta)'W(\theta - \delta)|\theta] \\ &= \mathbb{E}[\text{tr}(W(\theta - \delta)(\theta - \delta)')|\theta] \\ &= \mathbb{E}[\text{tr}(W(\theta - \mathbb{E}[\delta|\theta] + \mathbb{E}[\delta|\theta] - \delta)(\theta - \mathbb{E}[\delta|\theta] + \mathbb{E}[\delta|\theta] - \delta)')|\theta] \\ &= \text{tr}[W((\theta - \mathbb{E}[\delta|\theta])(\theta - \mathbb{E}[\delta|\theta])' + \mathbb{V}[\delta|\theta])] \; . \end{aligned}$$

This shows that unbiasedness is not necessarily a desirable property for a point estimator in the sense that there might be biased but less noisy estimators that yield a lower frequentist risk.

[5] The calculation can also be done for a sample of size n, but it is somewhat more involved. As usual, let $Y = (Y_1, ..., Y_n)'$, and consider $\delta = c'Y$, where c is an $n \times 1$ vector of weights, yielding the general point estimator $\delta = \sum_{i=1}^n c_i Y_i$. We know $Y \sim N(\iota\theta, I)$, where ι is a vector of ones. In turn, $c'Y \sim N(c'\iota\theta, c'Ic)$, and so $\mathbb{E}[(c'\iota)^2] = c'c + (c'\iota)^2\theta^2$. The frequentist risk is then

$$R(\theta, \delta) = \mathbb{E}[(\theta - c'Y)^2|\theta] = \theta^2 - 2\theta\mathbb{E}[c'Y|\theta] + \mathbb{E}[(c'Y)^2|\theta] = \theta^2(1 - c'\iota)^2 + c'c \; .$$

As $\iota'\iota = n$, this risk is minimized by $c = n^{-1}\iota$, i.e. by $\delta = \frac{1}{n}\sum_{i=1}^n y_i$.

2.1.2 Hypothesis Testing

Typically, we are not only interested in obtaining an estimator for θ, but also in testing whether θ lies in some set (interval) $\Theta_0 \subset \Theta$. More formally, we test the zero hypothesis $\mathscr{H}_0 = \theta \in \Theta_0$ against the alternative hypothesis $\mathscr{H}_1 : \theta \in \Theta_1$. Often, Θ_1 is simply the complement of Θ_0, i.e. all other regions of the parameter space: $\Theta_1 = \Theta_0^c = \Theta \backslash \Theta_0$. If Θ_0 contains a single value θ_0, then $\mathscr{H}_0 : \theta = \theta_0$ is called a point hypothesis, otherwise it is called a composite hypothesis (and analogously for $\mathscr{H}_1$). A hypothesis test can be viewed as a point estimator of the object $\mathbf{1}\{\theta \in \Theta_0\}$.

Definition 27. *A* hypothesis test $\varphi \in \{0,1\}$ *is a rule that specifies when we reject and when we accept (do not reject)* $\mathscr{H}_0$*, with* $\varphi = 0$ *indicating rejection.*

There are different tests that use different test-statistics. They are all constructed based on the idea that an event considered unlikely under $\mathscr{H}_0$ – such as an extreme realization – discredits $\mathscr{H}_0$. They all have the generic form

$$\varphi(y) = \mathbf{1}\{T(y) < c_\alpha\} \ , \tag{2.2}$$

i.e. we reject $\mathscr{H}_0$ if our realized test-statistic $T(y)$ is larger than some critical value c_α. To operationalize a hypothesis test φ, we need to choose a testing procedure, embodied by the test-statistic $T(y)$, and the critical value c_α.

The power function of a test is defined as the probability of rejecting $\mathscr{H}_0$ given that some $\theta \in \Theta$ is the true value:

$$\beta(\theta) = \mathbb{P}\left[\text{reject } \mathscr{H}_0 \mid \theta\right] = \mathbb{P}[\varphi = 0 \mid \theta] = \mathbb{P}\left[T(y) \geq c_\alpha \mid \theta\right] \ .$$

It is defined on the whole sample space Θ. A good test has a power function near 1 for $\theta \in \Theta_1$ and near 0 for $\theta \in \Theta_0$, i.e. a high probability of correct rejection (i.e. rejecting $\mathscr{H}_0$ if it is indeed false) and a low probability of false rejection (i.e. rejecting $\mathscr{H}_0$ if it is actually true). The probability of false rejection is referred to as the type I error, while the probability of wrong acceptance (i.e. one minus the probability of correct rejection) is referred to as the type II error.[6] Usually, we put a constraint on the type I error and then search for the test

[6] This terminology is linked to the fact how testing procedures φ are evaluated. This can be done using the following loss function:

$$L(\theta, \varphi) = \begin{cases} 1 & \text{if } \ \varphi \neq \mathbf{1}\{\theta \in \Theta_0\} & \text{(i.e. we are wrong)} \\ 0 & \text{otherwise} & \text{(i.e. we are right)} \end{cases} .$$

This loss function leads to the following frequentist risk $R(\theta, \varphi) = \mathbb{E}[L(\theta, \varphi(Y))|\theta]$ which depends on θ:

with the highest power.

Definition 28. *For $\alpha \in (0,1)$, a test is a* <u>*size α test*</u> *if $\sup\limits_{\theta \in \Theta_0} \beta(\theta) = \alpha$, i.e. if its highest possible type I error is equal to α.*[7]

Note that for a point hypothesis, the above expression simplifies to $\beta(\theta_0) = \alpha$. Different choices of the size α translate into different values for c_α in Eq. (2.2). Given α, the only other choice to make in order to fully specify a hypothesis test φ is to choose the test-statistic $T(Y)$, i.e. the testing procedure. As a first example of the latter, we consider the popular t-test.

Definition 29. *Suppose $\mathbb{E}[\hat{\theta}|\theta] = \theta$ and $\mathbb{V}[\hat{\theta}|\theta] = V$ and we are testing a point hypothesis $\mathscr{H}_0 : \theta = \theta_0$.*
Under the alternative $\mathscr{H}_1 : \theta \neq \theta_0$, the (two-sided) <u>*t-test*</u> *is $\varphi_t(y) = \mathbf{1}\left\{\left|\frac{\hat{\theta}-\theta_0}{v}\right| < c_\alpha\right\}$.*[8]

Note that the test-statistic $T(Y) = \left|\frac{\hat{\theta}-\theta_0}{v}\right|$ is a function of the data Y because our estimator of θ, $\hat{\theta}$, is a function of Y. Previously, we said that when we apply our estimator to our particular, observed realization of the data, y, we call the result an estimate. Similarly, we call $T(y)$ the realized test-statistic.

Let's take our running example and assume Normality: we have n observations $\{y_i\}_{i=1}^n$ of the RV $Y|\theta \sim N(\theta, 1)$. Section 2.1.1 showed that the LS/ML estimator $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n y_i \sim N\left(\theta, \frac{1}{n}\right)$.[9] If $\mathscr{H}_0$ is true, then this distribution is $N\left(\theta_0, \frac{1}{n}\right)$ and therefore $T(Y) = \frac{\hat{\theta}-\theta_0}{n^{-1/2}} \sim N(0,1)$. The two-sided t-test rejects $\mathscr{H}_0$ if this statistic is (on either side) too far from its supposed mean of zero (i.e. if $\hat{\theta}$ is too far from its supposed mean θ_0) and accepts otherwise.

- if $\theta \in \Theta_0$, then $R(\theta, \varphi) = \mathbb{P}\,[\text{reject } \mathscr{H}_0 \mid \mathscr{H}_0 \text{ is true}] \equiv a(\theta)$,
- if $\theta \in \Theta_1$, then $R(\theta, \varphi) = \mathbb{P}\,[\text{accept } \mathscr{H}_0 \mid \mathscr{H}_0 \text{ is false}] \equiv b(\theta)$.

Often, $a(\theta)$ is referred to as the type I error, $b(\theta)$ as the type II error, and $1 - b(\theta) = \mathbb{P}\,[\text{reject } \mathscr{H}_0 | \mathscr{H}_0 \text{ is false}]$ as the power of the test φ. However, note that $a(\theta)$ and $1 - b(\theta)$ are in fact the same function: the power function, which we write as $\beta(\theta)$.

[7] For discrete RVs, sometimes we cannot set $\sup\limits_{\theta \in \Theta_0} \beta(\theta)$ exactly to α. In these cases, we use the definition of a level α test; a test is a level α test if $\sup\limits_{\theta \in \Theta_0} \beta(\theta) \leqslant \alpha$, i.e. if the type I error is at most $\alpha \; \forall \; \theta \in \Theta_0$.

[8] Under the alternative $\mathscr{H}_1 : \theta > \theta_0$, the (one-sided) t-test is $\varphi_t(y) = \mathbf{1}\left\{\frac{\hat{\theta}-\theta_0}{v} > c_\alpha\right\}$, while under the alternative $\mathscr{H}_1 : \theta < \theta_0$, the (one-sided) t-test is $\varphi_t(y) = \mathbf{1}\left\{\frac{\hat{\theta}-\theta_0}{v} < c_\alpha\right\}$.

[9] In actual econometric models, $\hat{\theta}$ would be a more complicated function of the data. However, nothing changes conceptually; for example, under a linear regression model (see Chapter 3), we estimate a k-dimensional vector β, which (under certain assumptions) gives us $\hat{\beta}|\,\beta \sim N(\beta, V)$ and $\hat{\beta}_j|\beta_j \sim N(\beta_j, V_{jj})$ for a single element β_j of β.

To find c, we set the type I error equal to the desired size α:

$$\begin{aligned}
\alpha = \beta(\theta_0) &= \mathbb{P}[\varphi_t = 0|\theta = \theta_0] \\
&= 1 - \mathbb{P}[\varphi_t = 1|\theta = \theta_0] \\
&= 1 - \mathbb{P}\left[-c \leq \frac{\hat{\theta} - \theta_0}{n^{-1/2}} \leq c|\theta = \theta_0\right] \\
&= 1 - \mathbb{P}[-c \leq Z \leq c|\theta = \theta_0] \\
&= 1 - [\Phi(c) - \Phi(-c)] \\
&= 2(1 - \Phi(c)) \ ,
\end{aligned}$$

where $\Phi(c) = \mathbb{P}[Z \leq c]$ is the cdf of a standard Normal-distributed RV Z.[10] To get a test of size 10% ($\alpha = 0.1$), we take $c = 1.64$.

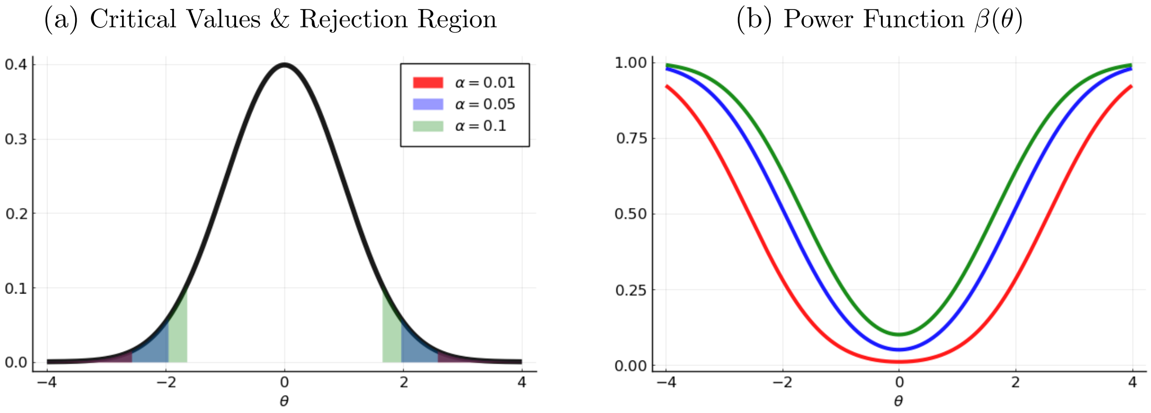
Above, we computed the power function β evaluated at θ_0. This is, by definition, the type I error, which we set equal to α. To derive the whole power function β, we need to consider not only $\theta \in \Theta_0$ (i.e. $\mathscr{H}_0$ being true), but also $\theta \in \Theta_1$, i.e. we need to consider the possibility that $\mathscr{H}_0$ might be false. To clarify this point, let us explicitly distinguish between the generic parameter θ, its supposedly true value under $\mathscr{H}_0$, θ_0, and its supposedly true value at which we evaluate the power function, $\tilde{\theta}$. It might be that $\tilde{\theta} = \theta_0$, i.e. $\mathscr{H}_0$ is true, but it might also be that $\tilde{\theta} \neq \theta_0$, i.e. $\mathscr{H}_0$ is false. If the true value of θ is $\tilde{\theta}$, then $\hat{\theta} \sim N\left(\tilde{\theta}, \frac{1}{n}\right)$ and so $T(Y) = \frac{\hat{\theta}-\theta_0}{n^{-1/2}} \sim N(\tilde{\theta} - \theta_0, 1)$. As a result,

$$\begin{aligned}
\beta(\tilde{\theta}) &= \mathbb{P}[\varphi_t = 0|\theta = \tilde{\theta}] \\
&= 1 - \mathbb{P}\left[-c \leq \frac{\hat{\theta} - \theta_0}{n^{-1/2}} \leq c|\theta = \tilde{\theta}\right] \\
&= 1 - \mathbb{P}[-c - (\tilde{\theta} - \theta_0) \leq Z \leq c - (\tilde{\theta} - \theta_0)|\ \theta = \tilde{\theta}] \\
&= 1 - [\Phi(c - (\tilde{\theta} - \theta_0)) - \Phi(-c - (\tilde{\theta} - \theta_0))] \ .
\end{aligned}$$

This function is plotted in Fig. 2.1. Different sizes of the test – embodied by different choices for c – lead to different powers of the test. The lower the size of the test – i.e. the smaller the probability of wrongly rejecting $\mathscr{H}_0$ – the lower also the power of the test – i.e. the higher the probability of wrongly accepting (or failing to reject) $\mathscr{H}_0$. It is common to be rather conservative (i.e. erring on the side of not rejecting $\mathscr{H}_0$) and report test results for sizes of 10%, 5% and 1%.[11]

[10] We used the fact that Φ is symmetric around zero: $\Phi(-c) = 1 - \Phi(c)$.

[11] Computing the power function of a test can become important in more complicated settings. Suppose we fail to reject some null hypothesis. This constitues evidence against an alternative hypothesis only if our


(a) Critical Values & Rejection Region

(b) Power Function $\beta(\theta)$

Figure 2.1: Two-Sided t-Tests

Notes: Illustration of the critical values, rejection region and power function for a two-sided t-test with sizes $\alpha \in \{0.01, 0.05, 0.1\}$.

To find the critical value and set up the test, we need to know the distribution of the test-statistic $T(y) = \frac{\hat{\theta}-\theta_0}{n^{-1/2}}$, which we know based on the distribution of $\hat{\theta}$: under $\mathscr{H}_0 : \theta = \theta_0$, $\hat{\theta} \sim N\left(\theta_0, \frac{1}{n}\right)$, and so we know $T(y) \sim N(0,1)$.[12] Recall from Section 2.1.1 that we know the distribution of our estimator $\hat{\theta}$ only if we assume the distribution of our data $Y|\theta$. Specifically, only when assuming that our data is Normal do we get that our estimator is Normal and hence also our test-statistic. However, regardless of the distribution of our data, our estimator is *asymptotically* Normal, which means that our test-statistic is asymptotically Normal and our testing procedure is asymptotically valid. Concretely, Section 2.1.1 showed that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0,1)$, i.e. $\hat{\theta} \overset{approx.}{\sim} N\left(\theta_0, \frac{1}{n}\right)$ in large samples, and $T(y) \xrightarrow{d} N(0,1)$.

Definition 30. *The* <u>*likelihood ratio (LR) test*</u> *is* $\varphi_{LR}(y) = \mathbf{1}\left\{\sup_{\theta\in\Theta_1} p(y|\theta) \Big/ \sup_{\theta\in\Theta_0} p(y|\theta) < c_\alpha\right\}$.

Intuitively, this ratio is small – and therefore the test is likely to accept $\mathscr{H}_0$ – if there are points in Θ_0 for which the observed y is much more likely than for points in Θ_1. For a point zero-hypothesis $\mathscr{H}_0 : \theta = \theta_0$ with the alternative $\mathscr{H}_1 : \theta \neq \theta_0$, the LR test simplifies to

$$\varphi_{LR}(y) = \mathbf{1}\left\{\frac{p(y|\hat{\theta}_{ML})}{p(y|\theta_0)} < c\right\} .^{13}$$

Consider our running example, $Y \sim N(\theta, 1)$, but suppose for simplicity we have a single

test has a (reasonably) high power under that alternative.

[12] More generally, under $\theta = \tilde{\theta}$, we have $\hat{\theta} \sim N\left(\tilde{\theta}, \frac{1}{n}\right)$ and so $T(y) \sim N(\tilde{\theta} - \theta_0, 1)$.

[13] This presumes that $\hat{\theta}_{ML} \in \Theta_1$, i.e. $\hat{\theta}_{ML} \neq \theta_0$. Under continuous parameter spaces like $\mathbb{R}$, the probability that $\hat{\theta}_{ML}$ is exactly equal to θ_0 is zero.

observation y of Y. We test $\mathscr{H}_0 : \theta = \theta_0$ against $\mathscr{H}_1 : \theta \neq \theta_0$. We get

$$T(y) = \frac{p(y|\hat{\theta}_{ML})}{p(y|\theta_0)} = \frac{(2\pi)^{-1/2} exp\left\{-\frac{1}{2}(y-\hat{\theta})^2\right\}}{(2\pi)^{-1/2} exp\left\{-\frac{1}{2}(y-\theta_0)^2\right\}} = exp\left\{\frac{1}{2}(\hat{\theta}-\theta_0)^2\right\} ,$$

as $\hat{\theta} = y$. To get a size α test, we find c_α so as to set the probability of wrong rejection to α:

$$\begin{aligned} \alpha = \beta(\theta_0) &= \mathbb{P}[\varphi_{LR} = 0|\theta = \theta_0] \\ &= 1 - \mathbb{P}[\varphi_{LR} = 1|\theta = \theta_0] \\ &= 1 - \mathbb{P}\left[exp\left\{\frac{1}{2}(\hat{\theta}-\theta_0)^2\right\} < c|\theta = \theta_0\right] \\ &= 1 - \mathbb{P}[(Y-\theta_0)^2 < 2\ln c|\theta = \theta_0] \\ &= F_{\chi^2_1}(2\ln c) , \end{aligned}$$

i.e. $2\ln c$ equals the $(1-\alpha)$th percentile of the Chi-squared distribution with one degree of freedom, due to the fact that $\hat{\theta} - \theta_0 \sim N(0,1)$ and therefore $(\hat{\theta}-\theta_0)^2 \sim \chi^2_1$.

Definition 31. *A test φ_α with size α and power function $\beta(\theta)$ is a uniformly most powerful size α test if it maximizes the power uniformly on Θ_1 among all tests with size α, i.e. if $\beta(\theta) \geqslant \beta'(\theta)$ for all $\theta \in \Theta_1$ and for all power functions $\beta'(\theta)$ of size α tests φ'.*

The Neyman-Pearson Lemma in the Appendix shows that if both $\mathscr{H}_0$ and $\mathscr{H}_1$ are point hypotheses, the LR test is uniformly most powerful.

Summing up, to set up a hypothesis test, we first choose a testing procedure like a t-test or an LR-test, and we then find the right critical value for our desired size of the test. Once this is done, we can finally apply the test to our particular dataset and the realized test-statistic it produces. Thereby, finding the critical value requires us to know the distribution of our test statistic $T(Y)$ under $\mathscr{H}_0$ (or a monotonic transformation of it, like in the previous example), because we need to know the probability of rejection under $\mathscr{H}_0$ – $\mathbb{P}[T(Y) > c|\mathscr{H}_0]$ – equal to α. We deduce this distribution based on the distribution of our estimator $\hat{\theta}|\theta$, which in turn is determined by the distribution of our data $Y|\theta$, all supposing that $\mathscr{H}_0$ is indeed true. Sometimes, however, we cannot proceed analytically. In these cases, we do the equivalent steps numerically: we numerically simulate the distribution of $T(Y)$ by repeatedly drawing data $Y|\theta$ from their supposed distribution under $\mathscr{H}_0$.

Consider the previous example, but with $\theta \in \mathbb{R}_+$ restricted to be positive. We get the ML estimator $\hat{\theta} = \max\{0, y\}$ and we can test $\mathscr{H}_0 : \theta = \theta_0$ vs. $\mathscr{H}_1 : \theta \neq \theta_0$ using the LR test with statistic

$$T(y) = 2\left[-\frac{1}{2}(x-\hat{\theta})^2 + \frac{1}{2}(x-\theta_0)^2\right] = (y-\theta_0)^2 - (y-\hat{\theta})^2 = \begin{cases} (y-\tilde{\theta})^2 & \text{if } y > 0 \\ \theta_0^2 - 2\theta_0 y & \text{if } y \leq 0 \end{cases} .$$

If $\mathcal{H}_0$ is true, we know $Y|\theta \sim N(\theta_0, 1)$, but we do not know the resulting distribution of the test statistic $T(Y)$. Instead, we can conduct our hypothesis test numerically as follows.

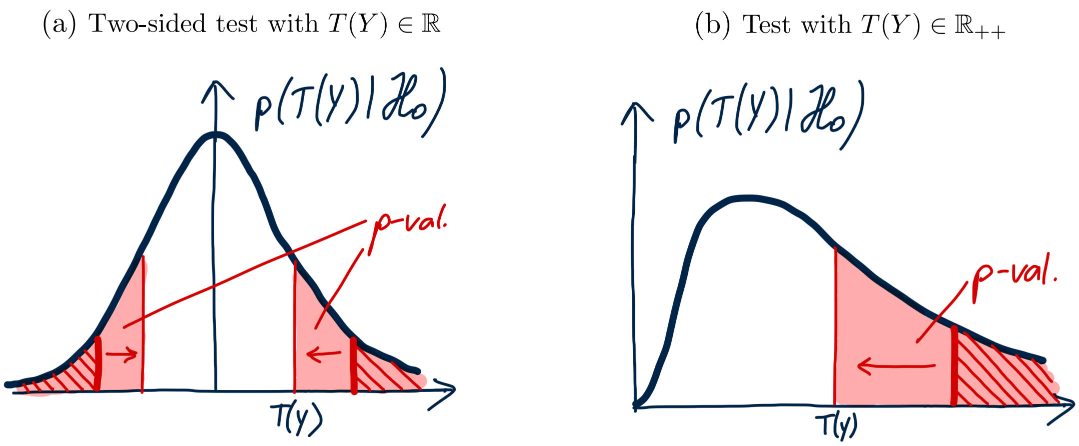
Algorithm 1 (Numerical Hypothesis Testing)**.**

1. *For $m = 1 : M$, draw y^m from $N(\theta_0, 1)$, the distribution of Y under $\mathcal{H}_0$. For each draw, compute $T(y^m)$. This gives you a distribution $\{T(y^m)\}_{m=1}^M$ that approximates the true distribution of $T(Y)$ under $\mathcal{H}_0$.*
2. *Sort $\{T(y^m)\}_{m=1}^M$ in ascending order and take the $M(1-\alpha)$th value as c_α. This empirical quantile approximates the $100(1-\alpha)$th quantile of the distribution of $T(Y)$ under $\mathcal{H}_0$.*
3. *Compute $T(y)$ for your particular, observed realization y. If $T(y) \leqslant c_\alpha$, then we accept, otherwise we reject.*

Definition 32. *A p-value is the largest type I error (i.e. largest size, probability of wrong rejection) of a test at which $\mathcal{H}_0$ is accepted (cannot be rejected).*
More formally, take a test $\varphi(y;\alpha)$ based on a test statistic $T(y)$ and let α be its size. Then p-value $= \sup\ \alpha$ s.t. $\varphi(y;\alpha) = 1$, i.e. we accept $\mathcal{H}_0$.

Recall the general structure of a test: $\varphi(y;\alpha) = \mathbf{1}\{T(y) < c_\alpha\}$, where we now emphasize the dependence of our test φ on the size α. Our examples illustrated that the size of a test is embodied in the critical value c based on which we judge our test-statistic $T(y)$, with larger sizes corresponding to smaller critical values. Of course, if we are conservative and take a very small size, i.e. small probability of wrong rejection, we have a large critical value, which means that we are likely to accept $\mathcal{H}_0$. In fact, as $\alpha \to 0$, we have $c \to \infty$, and we always accept! You can think of the p-value as the size obtained when we take the smallest c s.t. we accept $\mathcal{H}_0$. If the p-value is large, this means that we can even take a small c and still accept $\mathcal{H}_0$, which means that it is likely to be true. In contrast, if the p-value is small, this means that we indeed need a very large critical value to accept $\mathcal{H}_0$, which means that $\mathcal{H}_0$ is likely to be wrong (essentially we only accept it if we are very conservative, specifying a very small probability of wrong rejection). See illustration in Fig. 2.2.

Consider the two-sided t-test example from above: $Y \sim N(\theta, 1)$, and we test $\mathcal{H}_0 : \theta = \theta_0$ vs.


(a) Two-sided test with $T(Y) \in \mathbb{R}$

(b) Test with $T(Y) \in \mathbb{R}_{++}$

Figure 2.2: p-Values

Notes: Illustration of the critical values with corresponding rejection region (thick red line) and p-values (red area) for two different tests; left the two-sided t.test and right a test with positive support of the test-statistic (e.g. the LR test in the example in the text, where $T(Y) \sim \chi^2$ under $\mathscr{H}_0$).

$\mathscr{H}_1 : \theta \neq \theta_0$ using the test $\varphi(Y; \alpha) = \mathbf{1}\{T(Y) < c_\alpha\}$ based on the statistic $T(Y) = \left|\hat{\theta} - \theta_0\right|$, with $\hat{\theta} = Y$. Then

$$\begin{aligned} p &= \sup_\alpha \; \mathbb{P}\left[\varphi(Y;\alpha) = 0 | \mathscr{H}_0\right] \\ &= \sup_\alpha \; \mathbb{P}\left[|Y - \theta_0| > c_\alpha | \mathscr{H}_0\right] \\ &= \sup_\alpha \; 2 \cdot \mathbb{P}\left[Y - \theta_0 > c_\alpha | \mathscr{H}_0\right] \\ &= \sup_\alpha \; 2(1 - \Phi(c_\alpha)) \; . \end{aligned}$$

This expression decreases with c_α. Therefore, the smallest c_α we can take s.t. we still accept $\mathscr{H}_0$ is $c_\alpha = T(y) = |y - \theta_0|$. We obtain $p = 2(1 - \Phi(|y - \theta_0|))$.

The p-value reports the results of a test on a continuous $(0,1)$-scale rather than a discrete $\{0,1\}$-scale, i.e. accept/reject. It can be (and often is) used as a decision rule; we reject if the p-value is smaller than a desired size α, and accept otherwise.

2.1.3 Confidence Sets

In the frequentist paradigm, a confidence set $C(Y) \subseteq \Theta$ is a (random) set that should cover the true θ with a prespecified probability:

$$\inf_{\theta \in \Theta} \mathbb{P}[\theta \in C(Y) | \theta] = 1 - \alpha \; .$$

For a scalar θ, we speak of confidence intervals. (The infimum operator appears because generally, this probability can depend on the true value of θ, and because we do not know the true θ, the coverage probability has to be guaranteed $\forall\ \theta \in \Theta$). The probability is taken with respect to the RV Y, with θ fixed. Hence, the definition of $C(Y)$ above says that, if we look at many different, random realizations of Y, the set $C(Y)$ should contain the true θ in $100(1-\alpha)\%$ of cases (even if the true θ is such that this probability is lowest).

We can construct $C(Y)$ based on a test φ, testing the point-hypothesis $\mathscr{H}_0 : \theta = \theta_0$. Let $C(y)$ contain all the values for θ_0 that we would accept given our realization y:

$$C(y) = \{\theta_0 \in \Theta : \varphi(y;\theta_0) = 1\} \ .$$

Then

$$\inf_{\theta\in\Theta} \mathbb{P}[\theta \in C(Y)|\theta] = \inf_{\theta\in\Theta} \mathbb{P}[\varphi(y;\theta) = 1 \mid \theta] = 1 - \underbrace{\sup_{\theta\in\Theta} \left(\mathbb{P}[\varphi(y;\theta) = 0 \mid \theta]\right)}_{\text{type I error; reject } \theta \text{ given } \theta \text{ is true}} = 1-\alpha \ ,$$

i.e. we get the desired coverage probability.[14]

Consider the example from above: $Y \sim N(\theta, 1)$, and we test $\mathscr{H}_0 : \theta = \theta_0$ vs. $\mathscr{H}_1 : \theta \neq \theta_0$. Using the two-sided t-test $\varphi_t(y;\theta_0) = \mathbf{1}\{|y-\theta_0| < c^t_\alpha\}$, we accept $\mathscr{H}_0$ if $-c^t_\alpha \leqslant y-\theta_0 \leqslant c^t_\alpha$, which shows that the set of all θ_0 we would accept is $C^t_\alpha(y) = [y - c^t_\alpha, y + c^t_\alpha]$. In the case of a size $\alpha = 0.05$ test, we get the 95% confidence interval $C_{0.05,t}(y) = [y-1.96, y+1.96]$. Under the LR test, $\varphi_{LR}(y;\theta_0) = \mathbf{1}\left\{(y-\theta_0)^2 \leq c^{LR}_\alpha\right\}$, which means we accept $\mathscr{H}_0$ if $-\sqrt{c^{LR}_\alpha} \leq (y-\theta_0) \leq \sqrt{c^{LR}_\alpha}$ and yields $C^{LR}_\alpha(y) = [y - \sqrt{c^{LR}_\alpha}, y + \sqrt{c^{LR}_\alpha}]$.

For the same test statistic, there are (infinitely) many confidence sets one could construct. Typically, we want the confidence set to be as small (short) as possible (conditioning on a given coverage probability). Essentially, constructing $C(Y)$ from the acceptance region of a point-hypothesis test leads to the smallest $C(Y)$.

Just as sometimes, we cannot set up a test analytically, in these cases we cannot find an analytical expression for the confidence set, but have to construct it numerically. Consider the example from the previous section: $Y|\theta \sim N(\theta, 1)$, single realization y, with parameter space $\theta \in \mathbb{R}_+$. The previous section showed how to test $\mathscr{H}_0 : \theta = \theta_0$ vs. $\mathscr{H}_1 : \theta \neq \theta_0$ numerically using the LR test $\varphi(y;\theta_0) = \mathbf{1}\{T(y;\theta_0) < c\}$. Based on this test, we can

[14] Just as for some discrete distributions of Y, we cannot construct a test with size α, but with size at most α (i.e. $\leq \alpha$), for such distributions we cannot construct a $C(Y)$ for which the coverage probability is exactly $1-\alpha$. The actual, general definition of $C(Y)$ in the expression above uses a larger-or-equal sign: $\inf_{\theta\in\Theta} \mathbb{P}[\theta \in C(Y)|\theta] \geq 1-\alpha$.

construct $C(y)$ numerically as follows.

Algorithm 2 (Numerical Confidence Set Construction)**.**

1. *Choose a grid $\mathcal{T}$ of values for θ_0.*
2. *For each $\theta_0 \in \mathcal{T}$,*
   (a) *for $m = 1 : M$, draw y^m from $N(\theta_0, 1)$, the distribution of Y under $\mathcal{H}_0 : \theta = \theta_0$. For each draw, compute $T(y^m; \theta_0)$. This gives you a distribution $\{T(y^m; \theta_0)\}_{m=1}^M$ that approximates the true distribution of $T(Y; \theta_0)$ under $\mathcal{H}_0 : \theta = \theta_0$.*
   (b) *get the critical value for a size α test, $c_\alpha(\theta_0)$, as the (empirical) $100(1-\alpha)$th quantile of the distribution of $\{T(y^m; \theta_0)\}_{m=1}^M$.*
   (c) *compute $T(y; \theta_0)$ for your particular, observed realization y. If $T(y; \theta_0) \leqslant c_\alpha(\theta_0)$, then $\theta_0 \in C(y)$, otherwise $\theta_0 \notin C(y)$.*

2.2 Bayesian Inference

2.2.1 Point Estimation

The LS, MM and ML estimators all belong to the classical or frequentist paradigm, under which θ is thought to be an unknown, but fixed parameter. As opposed to that, in the Bayesian approach, θ is treated as a RV. Given some initial belief about its distribution (or about where the fixed, true value of θ lies) – the prior distribution – this belief is updated after observing the sample y to the posterior distribution (the distribution of θ or the belief about where the true value of θ lies after having observed the data) using Bayes' formula:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta) \; .$$

Bayesian inference often exploits the fact that the posterior is proportional to the product of the likelihood (i.e. the conditional distribution of y given θ) and the prior, which means that it (often) suffices to analyze the product of likelihood and prior in order to conclude something about the posterior (see Section 1.2). The term $p(y) = \int p(y|\theta)p(\theta)d\theta$ is just a normalization constant.

In our example, let our prior be $\theta \sim N(0, 1/\lambda)$, where λ is the inverse of the prior variance and specifies how confident we are in our belief that θ is equal to the prior mean of zero.

Under $Y|\theta \overset{i.i.d.}{\sim} N(0,1)$, we then get

$$\begin{aligned} p(\theta|y) &\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^2\right\}\exp\left\{-\frac{1}{2}\lambda\theta^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}(-2\theta\sum_{i=1}^{n}y_i+\theta^2(n+\lambda))\right\}\ , \end{aligned}$$

whereby we used the expression for $p(y|\theta)$ derived above. The above mentioned proportionality implies that we can drop all terms which do not depend on θ (i.e. all constants and all terms that depend only on y). From this expression, we can deduce that $\theta|y \sim N(\bar{\theta}, \bar{V})$ with $\bar{V} = \frac{1}{n+\lambda}$ and $\bar{\theta} = \frac{1}{n+\lambda}\sum_{i=1}^{n} y_i$.

Once the posterior $p(\theta|y)$ is obtained, there are several Bayesian point estimators one might construct (see next section). The most common ones are the posterior mean or median. In our example, both are equal to $\bar{\theta} = \frac{1}{n+\lambda}\sum_{i=1}^{n} y_i$. Note that this is a weighted average of the ML estimator and the prior mean, with weights given by the inverses of the ML estimator and prior variances, respectively:

$$\mathbb{E}[\theta|y] = \frac{\sum_{i=1}^{n} y_i}{n+\lambda} = \frac{1}{1/\mathbb{V}[\hat{\theta}_{ML}]+\lambda}\left[\frac{1}{\mathbb{V}[\hat{\theta}_{ML}]}\cdot\hat{\theta}_{ML}+\lambda\cdot 0\right]\ ,$$

whereby $\mathbb{V}[\hat{\theta}_{ML}] = n^{-1}$. If $\hat{\theta}_{ML}$ is precise (which happens if n is large, i.e. we have lots of data to draw inference from), the posterior mean will be closer to the ML estimator. The same happens if λ is low, which means that we are not very confident in our prior belief about θ. In contrast, as $\lambda \to \infty$, the posterior (mean) comes ever closer to the prior (mean).

Frequentist Properties of Bayesian Estimators Despite the fact that the posterior mean $\bar{\theta}$ is obtained in the Bayesian paradigm, we could analyze the properties of the point estimator $\hat{\theta}_B = \bar{\theta} = \frac{1}{n+\lambda}\sum_{i=1}^{n} Y_i$ just like the finite sample and asymptotic properties of the classical/frequentist estimators above. Note that this exercise is not a part of Bayesian inference, as it involves thinking about the properties of our estimator under different randomly drawn samples, whereas Bayesian inference conditions on the particular sample y obtained. However, estimators derived under the Bayesian paradigm are often used in classical/frequentist inference. For our point estimator $\hat{\theta}_B$, we can write $\hat{\theta}_B = \frac{n}{n+\lambda}\hat{\theta}_{ML}$ for $\hat{\theta}_{ML} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. In turn, we can see that $\hat{\theta}_B$ is biased,

$$\mathbb{E}[\hat{\theta}_B] = \frac{n}{n+\lambda}\mathbb{E}\left[\hat{\theta}_{ML}\right] = \frac{n}{n+\lambda}\theta \neq \theta\ ,$$

but consistent,

$$\hat{\theta}_B = \frac{n}{n+\lambda}\hat{\theta}_{ML} \xrightarrow{p} \theta \ ,$$

because $\hat{\theta}_{ML} \xrightarrow{p} \theta$ and $\lim_{n\to\infty} \frac{n}{n+\lambda} = 1$, and we can combine these two results using Slutsky's theorem. Also, we can show that

$$\mathbb{V}\left[\hat{\theta}_B\right] = \frac{n^2}{(n+\lambda)^2}\mathbb{V}[\hat{\theta}_{ML}] < \mathbb{V}[\hat{\theta}_{ML}] \quad \forall \quad \lambda > 0 \ .$$

Compared to $\hat{\theta}_{ML}$, $\hat{\theta}_B$ introduces a bias but reduces the variance. The intuition is that $\hat{\theta}_B$ is not based on the data alone, but shrinks $\hat{\theta}_{ML}$ to the prior mean of zero.

Evaluation of Estimators As introduced in Section 2.1.1, loss functions can be used to evaluate and compare point estimators. Under Bayesian inference, they are also used to motivate a particular point estimator $\hat{\theta}$ given a whole posterior distribution $p(\theta|y)$. Given a loss function $L(\theta, \delta)$, we can define Bayesian risk (or expected posterior loss).

Definition 33. *Bayesian risk:* $P(\mathbb{P}^\theta, \delta(y)) = \mathbb{E}[L(\theta, \delta(y))|y] = \int_\Theta L(\theta, \delta(y))p(\theta|y)d\theta$.[15]

Bayesian risk does the opposite of frequentist risk: it treats the data (the particular realization y we observe) as fixed and averages the loss function over all possible values of θ using the posterior distribution. As a result, in contrast to frequentist risk, Bayesian risk does not depend on θ, but on y, the realization of Y, which we know. Since Bayesian inference treats y as known, Bayesian risk yields a clear ordering of estimators. In fact, given the posterior $p(\theta|y)$, a Bayesian estimator is defined as the quantity that minimizes Bayesian risk.

Definition 34. *A Bayesian estimator (or Bayes estimator) associated with a prior distribution $\mathbb{P}^\theta$ and a loss function $L(\theta, \delta(Y))$ is an estimator that minimizes $P(\mathbb{P}^\theta, \delta(y))$.*

Under the quadratic loss function, the Bayes estimator is given by the posterior mean:

$$\min_{\delta\in\mathcal{D}} \mathbb{E}[(\theta - \delta(y))^2|y] \quad \Rightarrow \quad 2\mathbb{E}[\theta - \delta(y)|y] = 0 \quad \Rightarrow \quad \delta(y) = \mathbb{E}[\theta|y] \ .^{16}$$

Other loss functions can lead to other posterior moments as the Bayes estimator, e.g. the posterior median.

As somewhat of a sidenote, the definition of a prior distribution $p(\theta)$ also allows the compu-

[15] $\mathbb{P}^\theta$ denotes the prior distribution $p(\theta)$, but is written differently to avoid the wrong conclusion that Bayesian risk depends on θ.

[16] Note that $\mathbb{E}[(\theta - \delta(y))^2|y]$ must be bounded for this exchange of integration and differentiation to hold, i.e. the posterior variance must be bounded (see result in Appendix to Chapter 1).

tation of the so-called integrated risk. It computes the mean of the loss function using the joint distribution of θ and Y, i.e. it averages over all (a priori) possible parameter values as well as all possible realizations of Y. In contrast the frequentist risk, integrated risk gives a real number and therefore a way to compare point estimators that is independent of the true value of θ as well as of the data y.

Definition 35. *Integrated risk:* $r(\mathbb{P}^\theta, \delta) = \mathbb{E}[L(\theta, \delta(Y))] = \int_\mathcal{Y} \int_\Theta L(\theta, \delta(y))p(y, \theta)d\theta dy$.

Note that we can factorize $p(y, \theta)$ both as $p(y|\theta)p(\theta)$ and as $p(\theta|y)p(y)$. This means that we can re-write the integrated risk as in two different ways:

$$\begin{aligned} r(\mathbb{P}^\theta, \delta) &= \int_\mathcal{Y} \int_\Theta L(\theta, \delta(y))p(y, \theta)d\theta dy \\ &= \int_\Theta R(\theta, \delta)p(\theta)d\theta \\ &= \int_\mathcal{Y} P(\mathbb{P}^\theta, \delta(y))p(y)dy \ . \end{aligned}$$

Integrated risk is either the expectation of frequentist risk R taken w.r.t. the prior distribution $p(\theta)$, or the expectation of Bayesian risk P taken w.r.t. the marginal distribution of data $p(y)$. Since $p(y)$ is hard to obtain, it is easiest to compute integrated risk as $\mathbb{E}[R(\theta, \delta)]$, i.e. by first taking the conditional expectation using $p(Y|\theta)$ and then the expectation using $p(\theta)$. This calculation shows that minimizing the integrated risk means minimizing expected frequentist risk, whereby the expectation is taken using the prior distribution for θ.

2.2.2 Hypothesis Testing

As mentioned in Section 2.1.2, a hypothesis test can be viewed as a point estimator of the object $\mathbf{1}\{\theta \in \Theta_0\}$. The frequentist approach considers θ to be a fixed number. Therefore, $\mathbf{1}\{\theta \in \Theta_0\}$ is an unknown but fixed parameter. In other words, a hypothesis $\mathscr{H}_0 : \theta \in \Theta_0$ can either be true or false; θ either is in Θ_0 or it is not. In contrast, under the Bayesian approach, θ is a RV and we obtain probabilities for a hypothesis $\mathscr{H}_0$ (and $\mathscr{H}_1$) being true or false. The prior probability of $\mathscr{H}_0$ being true is $\mathbb{P}[\theta \in \Theta_0]$, while the posterior probability is $\mathbb{P}[\theta \in \Theta_0|Y]$. Hypotheses are then evaluated using

$$\text{posterior odds } = \frac{p(\theta \in \Theta_0|Y)}{p(\theta \in \Theta_1|Y)} \ ,$$

or

$$\text{Bayes factors } = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{p(\theta \in \Theta_0|Y)/p(\theta \in \Theta_1|Y)}{p(\theta \in \Theta_0)/p(\theta \in \Theta_1)} \ ,$$

The difference in testing philosophy between the two approaches is best illustrated using a simple example. Let $\Theta = \{0, 1\}$ and $\mathcal{X} = \{0, 1, 2, 3, 4\}$ with the following probabilities.

	0	1	2	3	4
$\mathbb{P}[Y\|\theta = 0]$	0.75	0.14	0.04	0.037	0.033
$\mathbb{P}[Y\|\theta = 1]$	0.7	0.251	0.04	0.005	0.004

Suppose we observe $y = 2$. As $P[Y \geqslant 2|\ \theta = 0] = 0.11$ and $\mathbb{P}[Y \geqslant 2|\theta = 1] = 0.049$, from the frequentist point of view, we would reject $\mathcal{H}_0 : \theta = 0$ at any $\alpha > 0.11$ and we would reject $\mathcal{H}_0 : \theta = 1$ at any $\alpha > 0.049$ (testing based on p-value). From the Bayesian point of view, if we consider $\theta = 1$ and $\theta = 0$ as equally likely a-priori, then $\mathbb{P}(\theta = j|y = 2) = \frac{\mathbb{P}(x=2|\theta=j)\mathbb{P}(\theta=j)}{\mathbb{P}(y=2)} = \frac{0.04 \cdot 0.5}{\mathbb{P}(y=2)}$ is the same for both $j = 0, 1$. We get the posterior odds $\frac{p(\theta=0|y=2)}{P(\theta=1|y=2)} = 1$ and so the observation $y = 2$ does not favor one model against the other. This illustrates that the Bayesian approach picks the better model (i.e. better value of θ) among all considered ones, while the Frequentist approach picks the conjectured model (value of θ) if it is "good" in the sense that the realization obtained is not too extreme/unlikely under the supposed model $\mathcal{H}_0$.

Two comments are in order. First, for Bayesian testing, there is no type I error; we cannot condition on $\mathcal{H}_0$ being true, because θ is not a parameter, but a RV in this perspective. Second, under continuous (prior and posterior) distributions for θ, non-trivial adjustments need to be made to test point hypotheses, because the probability of any single point is zero.

2.2.3 Credible Sets

Bayesian credible sets $C(Y)$ are defined by

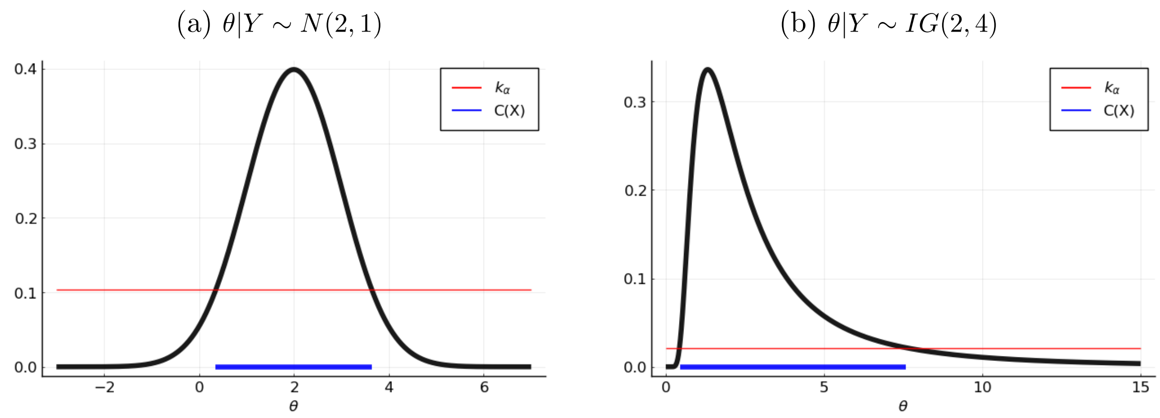
$$\mathbb{P}[\theta \in C(Y)|Y] = 1 - \alpha\ .$$[17]

In contrast to the frequentist paradigm, here X is fixed while θ is random. In line with the frequentist paradigm, here too there are many different sets $C(Y)$ one could construct. The smallest is the highest posterior density (HPD) set

$$C(Y) = \{\theta \in \Theta : p(\theta|Y) \geqslant k_\alpha\}\ , \quad \text{where} \quad k_\alpha \text{ is s.t. } \mathbb{P}[\theta \in C(Y)|\theta] \geqslant 1 - \alpha\ .$$

See illustration in Fig. 2.3. Essentially, we use a threshold value k_α to slice the posterior horizontally and look at the part of the y-axis (the values for θ) that correspond to the highest values of the posterior density.

[17] Again, more formally, this definition uses a greater-or-equal sign.


(a) $\theta|Y \sim N(2,1)$

(b) $\theta|Y \sim IG(2,4)$

Figure 2.3: Bayesian Highest Posterior Density (HPD) Sets

Notes: Illustration of the HPD sets with coverage probabilities $1-\alpha \in \{0.9, 0.95\}$ under different posteriors.

Appendix

Definition 36. *The minimax risk is* $\bar{R} = \inf_{\delta\in\mathcal{D}} \sup_{\theta\in\Theta} R(\theta,\delta)$*, where* $\mathcal{D}$ *is the set of all possible point estimators. The minimax estimator is* $\delta_0 = \arg\min_{\delta\in\mathcal{D}_0} (\max_{\theta\in\Theta} R(\theta,\delta))$.

The minimax estimator leads to the best among all possible worst cases. Intuitively, nature (by choosing θ) maximizes, the econometrician (by choosing δ) minimizes the risk.

Consider the example with $L = (\theta-\delta)^2$, $Y|\theta \sim N(0,1)$ and $\delta = c \cdot Y$. Let $\Theta = \mathbb{R}$. Taking $c = 1$ leads to the minimax estimator, since for any $c \neq 1$, nature could choose $\theta = \pm\infty$, which results in $R(\theta,\delta) = \pm\infty$ for any $\delta = cy$, $c \neq 1$.

Definition 37. *A point estimator* δ_1 *is inadmissible if* $\exists\ \delta_2$ *s.t.* $R(\theta,\delta_1) \geq R(\theta,\delta_2)\ \forall\ \theta$ *and* $\exists\ \theta'$ *s.t.* $R(\theta',\delta_1) > R(\theta',\delta_2)$.

In other words, δ_1 is inadmissible if there is another point estimator which never has a higher frequentist risk and for one value of θ even leads to a lower frequentist risk than δ_1. In our example, $\forall\ c > 1, R(\theta, cY) = \theta^2(1-c)^2 + c^2 > 1 = R(\theta, Y)$ and so any $c > 1$ is inadmissible.

Proposition 18 (Neyman-Pearson Lemma)**.**
Let $\mathcal{H}_0 : \theta = \theta_0$ *and* $\mathcal{H}_1 : \theta = \theta_1$*. Then* $\exists$ *a UMP test* $\forall\ \alpha \in (0,1)$ *and it is of the form* $\varphi(y) = \mathbf{1}\left\{\frac{p(y|\theta_1)}{p(y|\theta_0)} < k\right\}$*, with* k *determined so as to achieve the required type I error:* $\mathbb{P}\left[\frac{p(Y|\theta_1)}{p(Y|\theta_0)} > k|\theta_0\right] = \alpha$*. In other words, when both hypotheses are point hypotheses, the LR test is the UMP test.*

Proof: Take the above $\varphi(y)$. Suppose $\exists\ \varphi^*$ that satisfies the size constraint: $R(\theta_0, \varphi^*) = \int (1-\varphi^*) p(y|\theta_0) dy \leqslant \alpha$. If $y \in S^+$, $\varphi(y) = 1$ (we reject) and so $p(y|\theta_1) > kp(y|\theta_0)$. If $y \in S^-$, $\varphi(y) = 0$ (we do not reject) and so $p(y|\theta_1) < kp(y|\theta_0)$. Therefore,

$$\begin{aligned}
&\int (\varphi^r - \varphi)\left[p(y|\theta_1) - kp(y|\theta_0)\right] dy \\
&\quad = \int_{S^+} (\varphi^* - \varphi)\left[p(y|\theta_1) - kp(y|\theta_0)\right] dy + \int_{S^-} (\varphi^* - \varphi)\left[p(y|\theta_1) - kp(y|\theta_0)\right] dy \quad \geqslant 0\ .
\end{aligned}$$

This means that

$$\begin{aligned}
\beta(\theta_1) - \beta^*(\theta_1) &= \int [(1-\varphi) - (1-\varphi^*)]\rho(y|\theta_1) dy \\
&= \int (\varphi^* - \varphi) p(y|\theta_0) dy \\
&\geqslant \int (\varphi^* - \varphi) k\rho(y|\theta_0) dy \\
&= k[1 - \alpha^* - (1-\alpha)] \\
&= k(\alpha - \alpha^*) \geqslant 0
\end{aligned}$$

as $\alpha^* \leqslant \alpha$. ■