LECTURE 11 NOTES

A hypothesis is a statement regarding the distribution of a random variable $x \in \mathcal{X}$, or, in a parametric setup, a parameter θ. Hypotheses occur in pairs: a *null hypothesis* of the form

$$H_0 : \theta \in \Theta_0 \subset \Theta$$

and an *alternative hypothesis* of the form

$$H_1 : \theta \in \Theta_1 := \Theta \setminus \Theta_0.$$

The goal of hypothesis testing is to decide between the null and alternative hypotheses based on observations $\mathbf{x} \sim F_\theta$.

Formally, a statistical test of H_0 is a binary statistic $\varphi : \mathcal{X} \to \{0, 1\}$. By convention, $\varphi(\mathbf{x}) = 0$ corresponds to accepting H_0, and $\varphi(\mathbf{x}) = 1$, rejecting H_0. Thus φ partitions the sample space into two regions:

$$\mathcal{X} = \underbrace{\{x \in \mathcal{X} : \varphi(x) = 0\}}_{\text{acceptance region}} \bigcup \underbrace{\{x \in \mathcal{X} : \varphi(x) = 1\}}_{\text{critical/rejection region}}.$$

In other words, the test φ is simply the indicator function of the critical region. Thus specifying a critical region is equivalent to specifying a test. The expected value of φ as a function of θ is called the *power function*:

$$\beta(\theta) := \mathbf{E}_\theta\big[\varphi(\mathbf{x})\big] = \mathbf{P}_\theta\big(\varphi(x) = 1\big).$$

It is the chance of rejecting H_0 as a function of θ. The power function of a “good” test is small on Θ_0, and large (close to one) on Θ_1.

Typically, a test φ has the form

$$\varphi(\mathbf{x}) = \mathbf{1}_{\mathcal{C}}(\phi(\mathbf{x})),$$

where $\phi(\mathbf{x})$ is called a *test statistic* and $\mathcal{C} \subset \phi(\mathcal{X})$ is the critical region. The preceding test accepts or rejects the null depending only on the test statistic $\phi(\mathbf{x})$. For example, to decide whether the mean of a Gaussian random variable is positive, a test may reject H_0 if the sample mean is smaller than -1.96. Here the test statistic is the sample mean.

In hypothesis testing, the action space is binary. Thus the test may only commit one of two types of errors:

1. *Type I errors:* reject H_0 when it is true,
2. *Type II errors:* accept H_0 when it is false.

The worst-case probability of committing a Type I error is the *significance level* of a test:

$$\sup_{\theta\in\Theta_0}\beta(\theta) = \sup_{\theta\in\Theta_0}\mathbf{E}_\theta\big[\varphi(\mathbf{x})\big].$$

The probability of rejecting H_0 when it is false is the *power* of the test. We see that

$$\beta(\theta) = \begin{cases}\text{Type I error rate} & \theta\in\Theta_0\\ 1-\text{Type II error rate} & \theta\in\Theta_1.\end{cases}$$

By convention, the null hypothesis represents the status quo. It is the statement that the investigator defaults to in the absence compelling evidence to the contrary. Thus Type I errors are also called *false discoveries.* Type I errors are considered more serious than Type II errors, and attention is usually focused on tests that control the Type I error rate. In the classical approach to hypothesis testing, the investigator specifies a tolerance α for Type I errors and considers only α-level tests.

Sometimes, it is not possible to derive an exact α-level test; i.e. a test φ such that

$$\mathbf{P}_0(\varphi(\mathbf{x}) = 1) = \alpha,$$

especially when $\mathbf{x}$ is a discrete random variable. The work-around is to consider *randomized tests*; i.e. tests that, for some observations $x\in\mathcal{X}$, rejects H_0 with some judiciously chosen probability, so that the overall probability of rejecting H_0 is exactly α under H_0. To avoid the technicalities that arise when evaluating randomized tests, we assume it is possible to derive an exact α-level test.

1. The likelihood ratio test. The likelihood ratio test (LRT) is the counterpart of the MLE in hypothesis testing. It is also based on the likelihood function and thus broadly applicable.

Definition 1.1. *The likelihood ratio test for testing $H_0:\theta\in\Theta_0$ versus $H_1:\theta\in\Theta_1 := \Theta\setminus\Theta_0$ is $\varphi_{\mathrm{LRT}}(x) := \mathbf{1}_{(t,\infty)}(\lambda(x))$, where*

$$\lambda(x) := \frac{\sup_{\theta\in\Theta} l_x(\theta)}{\sup_{\theta\in\Theta_0} l_x(\theta)}$$

is the likelihood ratio statistic.

Since $\Theta_0\subset\Theta$, the likelihood ratio is at least one. If H_0 is true, the MLE of θ is likely close to Θ_0, and we expect $\lambda(\mathbf{x})$ to be close to one. If H_0 is false, we expect $\lambda(\mathbf{x})$ to be considerably larger than one. Thus we reject H_0 when $\lambda(\mathbf{x})$ is large.

The critical value t controls the level of the test. A larger t is more *conservative*: the test is less likely to reject H_0. A smaller critical value is more powerful, but also increases the Type I error rate. In the classical approach to hypothesis testing, investigators set t as small as possible (to maximize power) while keeping the Type I error rate below α. Occasionally, it is possible to set t exactly.

EXAMPLE 1.2. *Let* $\mathbf{x} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$. *Consider testing* $H_0 : \mu = \mu_0$ *versus* $H_1 : \mu \neq \mu_0$. *It is possible to show that the (unconstrained) MLE of* μ *is* $\bar{\mathbf{x}}$. *Thus the likelihood ratio is*

$$\begin{aligned}\lambda(\mathbf{x}) &= \frac{(2\pi)^{-n/2}\prod_{i\in[n]}\exp\left(-\frac{1}{2}(\mathbf{x}_i - \bar{\mathbf{x}})^2\right)}{(2\pi)^{-n/2}\prod_{i\in[n]}\exp\left(-\frac{1}{2}(\mathbf{x}_i - \mu_0)^2\right)} \\ &= \exp\left(\tfrac{1}{2}\textstyle\sum_{i\in[n]}(\mathbf{x}_i - \mu_0)^2 - \tfrac{1}{2}\sum_{i\in[n]}(\mathbf{x}_i - \bar{\mathbf{x}})^2\right) \\ &= \exp\left(\tfrac{n}{2}(\bar{\mathbf{x}} - \mu_0)^2\right).\end{aligned}$$

The probability of $\lambda(\mathbf{x}) > t$ *is the same as the probability of*

$$2\log\lambda(\mathbf{x}) = n(\bar{\mathbf{x}} - \mu_0)^2 > 2\log t.$$

To obtain an α*-level test, we should set* t *so that*

$$\mathbf{P}_0\big(n(\bar{\mathbf{x}} - \mu_0)^2 > 2\log t\big) \leq \alpha \tag{1.1}$$

Since $n(\bar{\mathbf{x}} - \mu_0)^2$ *is a* χ_1^2 *random variable, we should set* t *so that* $2\log t = \chi_{1,\alpha}^2$, *where* $\chi_{1,\alpha}^2$ *is the* $1 - \alpha$ *quantile of a* χ_1^2 *random variable. The power function of the preceding test is*

$$\begin{aligned}\beta(\mu) &= \mathbf{P}_\mu\big(n(\bar{\mathbf{x}}_n - \mu_0)^2 > \chi_{1,\alpha}^2\big) \\ &= \mathbf{P}_\mu\big(\big|\sqrt{n}(\bar{\mathbf{x}}_n - \mu_0)\big| > z_{\frac{\alpha}{2}}\big) \\ &= \mathbf{P}_\mu\big(\big|\sqrt{n}(\bar{\mathbf{x}}_n - \mu) + \sqrt{n}(\mu - \mu_0)\big| > z_{\frac{\alpha}{2}}\big) \\ &= \mathbf{P}\big(\big|\mathbf{z} + \sqrt{n}(\mu - \mu_0)\big| > z_{\frac{\alpha}{2}}\big),\end{aligned}$$

where $\mathbf{z}$ *is a standard normal random variable.*

In the design of statistical investigations, a key parameter is the sample size required to reliably reject a given alternative. For example, how many samples are required to achieve 0.9 power against the alternative $H_1 : \mu = \mu_0 + 0.1$? *The requisite sample size is given by the solution of*

$$\beta(\mu) = \mathbf{P}\big(\big|\mathbf{z} + 0.1\sqrt{n}\big| > z_{\frac{\alpha}{2}}\big) \geq 0.9.$$

The power it at least

$$\begin{aligned}\beta(\mu) &= \mathbf{P}\big(\mathbf{z} + 0.1\sqrt{n} < -z_{\frac{\alpha}{2}}\big) + \mathbf{P}\big(\mathbf{z} + 0.1n > z_{\frac{\alpha}{2}}\big)\\ &\ge \mathbf{P}\big(\mathbf{z} + 0.1\sqrt{n} > z_{\frac{\alpha}{2}}\big)\\ &= 1 - \Phi\big(z_{\frac{\alpha}{2}} - 0.1\sqrt{n}\big).\end{aligned}$$

We solve for n to obtain $n > 325$.

More often than not, it is not possible to set t exactly. However, when H_0 is a *simple hypothesis* (Θ_0 is a singleton), it is possible to simulate the distribution of $\lambda(\mathbf{x})$ under H_0.

Example 1.3. *Let $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p)$. Consider testing $H_0 : p = \frac{1}{2}$ versus $H_1 : p \ne \frac{1}{2}$. We showed that the MLE of p is $\frac{\mathbf{t}}{n}$, where $\mathbf{t} = \sum_{i\in[n]} \mathbf{x}_i$. The likelihood ratio is*

$$\lambda(\mathbf{t}) = \frac{\prod_{i\in[n]} \hat{p}^{\mathbf{x}_i} (1-\hat{p})^{1-\mathbf{x}_i}}{\left(\frac{1}{2}\right)^n} = \left(\tfrac{1}{2}\right)^{-n} \left(\tfrac{\mathbf{t}}{n}\right)^{\mathbf{t}} \left(1 - \left(\tfrac{\mathbf{t}}{n}\right)\right)^{n-\mathbf{t}},$$

It is not possible to set t by evaluating $\mathbf{P}(\lambda(\mathbf{x}) > t)$ explicitly, but it is possible to evaluate the CDF of $\lambda(\mathbf{x})$ under H_0. For numerical stability, we simulate the distribution of

$$\log\lambda(\mathbf{t}) := \mathbf{t}\log\big(\tfrac{\mathbf{t}}{n-\mathbf{t}}\big) + n\log\big(1 - \tfrac{\mathbf{t}}{n}\big) - n\log\tfrac{1}{2}.$$

To ensure $\mathbf{P}_0(\lambda(\mathbf{t}) > t)$ is at most α, we set t to be the $1-\alpha$ quantile of the simulated distribution.

We observe that in the preceding examples, the LR depend on the observations only through sufficient statistics. This is no mere coincidence. If $\mathbf{t} = \phi(\mathbf{x})$ is a sufficient statistics, by the factorization theorem,

$$\lambda(x) = \frac{\sup_{\theta\in\Theta} f_\theta(x)}{\sup_{\theta\in\Theta_0} f_\theta(x)} = \frac{\sup_{\theta\in\Theta} g_\theta(\phi(x))}{\sup_{\theta\in\Theta_0} g_\theta(\phi(x))}.$$

When H_0 is a *composite hypothesis* (Θ_0 is not a singleton), it is usually hard to set the rejection threshold exactly, and we must resort to asymptotic approximations. Composite hypotheses occur when there are *nuisance parameters*, i.e. parameters that are part of the model but not the target of investigation. Since the investigator is not concerned with nuisance parameters, they are allowed to take any value, leading to composite hypotheses of the form

$$\Theta_0 = \{(\theta_1, \theta_2) \in \Theta : \theta_1 = 0\}. \tag{1.2}$$

Example 1.4. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$*, where* σ^2 *is unknown. Consider testing* $H_0 : \mu = \mu_0$ *versus* $H_1 : \mu \neq \mu_0$*. It is possible to show that*

1. *the (unconstrained) MLE of* μ *is* $\bar{\mathbf{x}}$
2. *the MLE of* σ *is* $\hat{\sigma}^2 = \frac{1}{n}\sum_{i\in[n]}(\mathbf{x}_i - \bar{\mathbf{x}})^2$
3. *the constrained MLE of* σ^2 *is* $\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i\in[n]}(\mathbf{x}_i - \mu_0)^2$.

Thus the likelihood ratio is

$$\begin{aligned}\lambda(\mathbf{x}) &= \frac{(2\pi\hat{\sigma}^2)^{-n/2}\exp\left(-\frac{1}{2\hat{\sigma}^2}\sum_{i\in[n]}(\mathbf{x}_i - \bar{\mathbf{x}})^2\right)}{(2\pi\hat{\sigma}_0^2)^{-n/2}\exp\left(-\frac{1}{2\hat{\sigma}_0^2}\sum_{i\in[n]}(\mathbf{x}_i - \mu_0)^2\right)} \\ &= \frac{\hat{\sigma}_0^n e^{-\frac{n}{2}}}{\hat{\sigma}^n e^{-\frac{n}{2}}} = \frac{\hat{\sigma}_0^n}{\hat{\sigma}^n}.\end{aligned}$$

The event $\lambda(\mathbf{x}) > t$ *is equivalent to*

$$t^{\frac{1}{n}} < \frac{\sum_{i\in[n]}(\mathbf{x}_i - \mu_0)^2}{\sum_{i\in[n]}(\mathbf{x}_i - \bar{\mathbf{x}})^2} = 1 + \frac{n(\bar{\mathbf{x}} - \mu_0)^2}{\sum_{i\in[n]}(\mathbf{x}_i - \bar{\mathbf{x}})^2},$$

which is in turn equivalent to

$$\left|\frac{\sqrt{n}(\bar{\mathbf{x}} - \mu_0)}{\left(\sum_{i\in[n]}(\mathbf{x}_i - \bar{\mathbf{x}})^2\right)^{1/2}}\right| > \left(t^{\frac{1}{n}} - 1\right)^{\frac{1}{2}}.$$

We multiply by $(n-1)^{\frac{1}{2}}$ *to obtain* $|\phi(\mathbf{x})| > \left(\left(t^{\frac{1}{n}} - 1\right)(n-1)\right)^{\frac{1}{2}}$ *where*

$$\phi(\mathbf{x}) := \frac{\sqrt{n}(\bar{\mathbf{x}} - \mu_0)}{\left(\frac{1}{n-1}\sum_{i\in[n]}(\mathbf{x}_i - \bar{\mathbf{x}})^2\right)^{1/2}}$$

is the t-statistic. To obtain an α*-level test, we reject* H_0 *when*

$$|\phi(\mathbf{x})| > t_{n-1,\frac{\alpha}{2}}.$$

Figure 1 plots the power functions of the t-test and z-test. The t-test is less powerful than the z-test, which is the price we pay for not knowing σ^2*.*

Thus far, we chose critical regions intuitively: the LR is large under the alternative, so we should reject the null when the LR is large. However, rejecting when the LR is smaller than its α-quantile is another α-level test. As we shall see, the intuitive choice is also the most powerful choice.

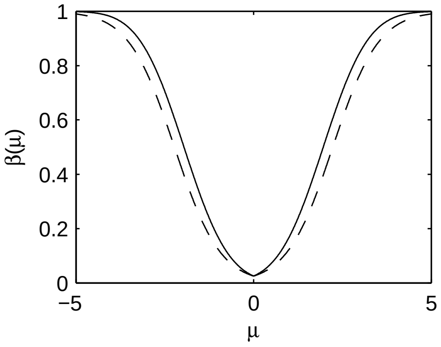

Fig 1: The power of the LRT for testing $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ when σ^2 is known (solid curve) and when σ^2 is unknown (dotted curve)

2. p-values.

Definition 2.1. *A* p-value *is a statistic* $p : \mathcal{X} \to [0,1]$ *that is stochastically larger than* unif$(0,1)$ *under the* H_0*:*

$$\inf_{\theta\in\Theta_0} \mathbf{P}_\theta(p(\mathbf{x}) > \alpha) > 1 - \alpha \text{ for any } \alpha \in [0,1]. \tag{2.1}$$

It is possible to check that (2.1) is equivalent to

$$\sup_{\theta\in\Theta_0} \mathbf{P}_\theta(p(\mathbf{x}) \le t) \le t \text{ for any } t \in [0,1].$$

The requirement that a p-value is stochastically larger than unif$(0,1)$ under H_0 is convention. By convention, practitioners concoct p-values so that small values are grounds for rejecting H_0. To avoid false rejections, pvalues are required to the (stochastically) large under H_0.

The easiest way to obtain a p-value from a test statistic $\phi(\mathbf{x}) \in \mathbf{R}$ that is stochastically larger under the alternative (e.g. the LR statistic) is

$$p(\mathbf{x}) := \sup_{\theta\in\Theta_0} 1 - F_\theta(\phi(\mathbf{x})),$$

where $F_\theta(t) := \mathbf{P}_\theta\big(\phi(\mathbf{x}) \le t\big)$ is the CDF of $\phi(\mathbf{x})$. That is, $p(\mathbf{x})$ is one less the CDF $\phi(\mathbf{x})$ evaluated at the observed value of $\phi(\mathbf{x})$. Indeed,

$$\begin{aligned}\mathbf{P}_{\theta_0}\big(p(\mathbf{x}) \le \alpha\big) &= \mathbf{P}_{\theta_0}\big(\sup_{\theta\in\Theta_0} 1 - F_\theta(\phi(\mathbf{x})) \le \alpha\big) \\ &\le \mathbf{P}_{\theta_0}\big(1 - F_{\theta_0}(\phi(\mathbf{x})) \le \alpha\big) \\ &= \mathbf{P}_{\theta_0}(\phi(\mathbf{x}) \ge F_{\theta_0}^{-1}(1-\alpha)) \\ &= \alpha.\end{aligned}$$

The preceding p-value is the probability under the null that $\phi(\mathbf{x})$ is larger than the observed value t. Intuitively, it is the probability of observing a test statistic that is even more extreme than the observed valued of the test statistic.

Conversely, as long as $p(\mathbf{x})$ is a p-value (i.e. $p(\mathbf{x})$ is stochastically smaller than uniform under the null),

$$\varphi(\mathbf{x}) := \mathbf{1}_{[0,\alpha]}(p(\mathbf{x}))$$

is a α-level test of the null. Indeed, for any $\theta \in \Theta_0$

$$\mathbf{E}_\theta\big[\varphi(\mathbf{x})\big] = \mathbf{P}_\theta\big(p(\mathbf{x}) \le \alpha\big),$$

which is at most α by Definition 2.1. Thus $\varphi(\mathbf{x})$ is an α-level test.

Intuitively, a p-value is a summary of the evidence in the observation $\mathbf{x}$ against the null: since small p-values are more likely under the alternative, a small p-value is interpreted as evidence agains the null. Thus reporting a p-value is a "continuous" summary of the conclusion of a hypothesis test whereas just reporting the outcome is a coarse binary summary.

We hasten to remark that observing $p(\mathbf{x}) = p$ should not be interpreted as "the null is true with probability p". The null is either true or false: it is non-random. Hence it is nonsense to speak of the probability of the null being true. A correct statement is "the probability (under the null) of observing a p-value smaller than p is at most p".

Example 2.2 (Example 1.2 continued). *A p-value for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ is $p(\mathbf{x}) = 1 - F_{\chi_1^2}\big((\mathbf{x} - \mu_0)^2\big)$.*

Example 2.3 (Example 1.3 continued). *An approximate pvalue for testing $H_0 : p = \frac{1}{2}$ versus $H_1 : p \neq \frac{1}{2}$ is*

$$p(\mathbf{x}) = 1 - \widehat{F}_0\big(\log \lambda(\mathbf{t})\big),$$

where $\widehat{F}_0$ is the simulated CDF of $\log \lambda(\mathbf{t})$.

Recently, the use (or misuse) of p-values in science has stirred up controversy in the scientific community. At the end of the day, p-values are just statistics, and their interpretation is ultimately up to the investigator and the broader scientific community. To quote Ioannidis (2005),

> It is misleading to emphasize the statistically significant findings of any single team. What matters is the totality of the evidence.

As long as the p-values are correctly interpreted, it is perfectly sound to support research findings by reporting p-values.

References.

IOANNIDIS, J. P. (2005). Why most published research findings are false. *Chance* **18** 40–47.

YUEKAI SUN
BERKELEY, CALIFORNIA
DECEMBER 2, 2015