9.1 Framework for Testing Hypotheses

Hypotheses and Power

Suppose I toss a coin 10 times, and then truthfully tell you that all 10 flips were heads. What are the possible explanations?

1. **Chance**: The coin is fair, the tosses were independent, but we observed a rare but not impossible even. (We should see this result about once in every $2^{10} \approx 1000$ flips).
2. **Bias**: The coin isn't fair, but rather, the coin has a high probability of landing heads (closer to 1 than $\frac{1}{2}$.
3. **Dependence**: The coin flips aren't actually independent. Maybe I got lazy and just flipped the coin once and wrote down the result 9 times.

Which of these outcomes is most likely? *It's impossible to say without assigning prior probabilities.*

However, we could ask: for which of these possibilities was the data most consistent? In this case, the observed event would be considered rare if the coin flips were independent and unbiased. But it would be rather mundane if the coin is either biased towards heads, or the flips were dependent.

When choosing among possible hypotheses, we gravitate towards those for which the observed data was commonplace. (Rightly or wrongly) This is the framework for the hypothesis test.

Def: Let $\mathcal{P}$ be a statistical model with parameter space Ω, and let Ω_1, Ω_2 be a partition of Ω. The **null hypothesis** is the claim $H_0 : \theta \in \Omega_0$ and the **alternative hypothesis** is the claim $H_1 : \theta \in \Omega_1$. If Ω_i consists of a single value, we say that H_i is a **simple hypothesis**. Any hypothesis that is not simple is a **composite hypothesis**. In the special case when H_i is an interval of the form (a, ∞) or $(-\infty, a)$, then we say it is a **one-sided hypothesis**. And when H_i is of the form $(-\infty, a) \cup (a, \infty)$ we say that H_i is a **two-sided hypothesis**.

Ex 1: Suppose we want to test whether a coin is fair. Our model assumes the sample $\mathbf{X}$ are conditionally independent Bernoulli-θ variables, with parameter space $\Omega = [0, 1]$. Our hypotheses are

$$H_0 : \theta = 0.5 \qquad H_a : \theta \neq 0.5$$

That is, $\Omega_0 = \{0.5\}$ and $\Omega_1 = [0, 0.5) \cup (0.5, 1]$.

Our ultimate goal is to use data to draw a conclusion about the plausibility of H_0 and H_1, gravitating towards the hypothesis for which the data is most consistent.

Ex 2: Continuing with the coin flip example above, we might flip a coin 10 times and compute the number of heads $\sum X_i$ observed. If $\sum X_i$ is far from 5, it seems reasonable to reject H_0 in favor of H_1. In particular, perhaps we decide to use the rule "if $|\sum X_i - 5| > 3$, then we will reject H_0."

Note that this divides our sample space S into two disjoint regions:

$$S_0 = \{\mathbf{x} \in S : \left|\sum x_i - 5\right| \leq 3\} \qquad S_1 = \{\mathbf{x} \in S : \left|\sum x_i - 5\right| > 3\}$$

That is, S_0 consists of all observed data $\mathbf{x}$ where we do not reject H_0 and where S_1 consists of observed data where we do reject H_0.

Def: A **Hypothesis Test Procedure** δ is a specification of:

- a parameterized model $\mathcal{P}$ for the data;
- hypotheses H_0 and H_1 which partition the parameter space Ω into two parts Ω_0 and Ω_1;

- a partition of the sample space S into two parts S_0 and S_1, where S_1 is called the **critical region** and represents the samples for which we reject H_0.

In many cases, we will just use the values of a particular statistic $T(\mathbf{X})$, rather than the entire sample data $\mathbf{X}$, to determine the partition of the sample space S. In this case, we call $T(\mathbf{X})$ a **test statistic**, and call $R = T(S_1) \subset \mathrm{R}$ the **rejection region**.

Ex 3: Using the coin flip model, let $T(\mathbf{X}) = |\sum X_i - 5|$. With $n = 10$, we might decide that $R = (3, \infty)$. That is, we reject H_0 if $T > 3$.

Once a hypothesis procedure δ has been specified, we can investigate the probability that the test rejects H_0.

Def: Let δ be a test procedure. Define $\pi(\theta|\delta)$ to be the function of θ given by

$$\pi(\theta|\delta) = P(\mathbf{X} \in S_1 \,|\, \theta)$$

That is, $\pi(\theta|\delta)$ is the probability of rejecting H_0, if the true value of the parameter is θ. We say that $\pi(\theta|\delta)$ is the **power function** of the test δ.

An ideal power function would be one with

$$\pi(\theta|\delta) = \begin{cases} 1, & \theta \in \Omega_1, \\ 0, & \theta \in \Omega_0 \end{cases}$$

That is, a power function corresponding to a test that always rejects H_0 when it is false, and that never rejects H_0 when it is true. In practice, however, we will never have power functions with this property (otherwise, a sample of data would give us perfect information about the parameter).

In general, we can make two types of errors when performing a hypothesis test:

	H_0 true	H_1 true
Reject H_0	type I error	Correct
Do not Reject H_0	Correct	type II error

If $\theta \in \Omega_0$, the power function is the probability of a type I error. While if $\theta \in \Omega_1$, the probability of a type II error is $1 - \pi(\theta|\delta)$.

All else equal, we want a test procedure for which $\pi(\theta|\delta)$ is close to 0 for $\theta \in \Omega_0$ and $\pi(\theta|\delta)$ is close to 1 for $\theta \in \Omega_1$.

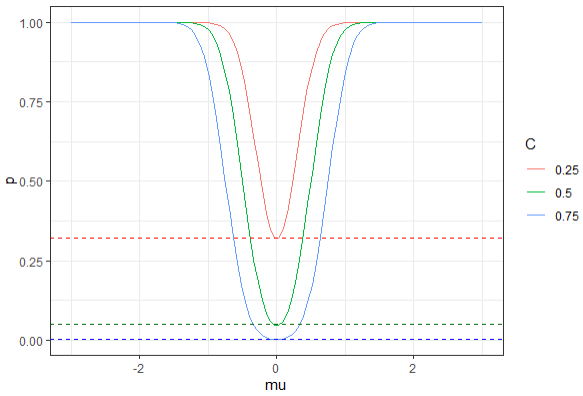
Ex 4: Consider the naive test procedure which is $S_1 = S$. Then

$$\pi(\theta|\delta) = \begin{cases} 1, & \theta \in \Omega_0 \\ 0, & \theta \in \Omega_1 \end{cases}$$

Ex 5: Suppose a sample $\mathbf{X}$ of size n is drawn from a Normal population with unknown mean μ and known variance σ^2 and consider hypotheses $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$. Let $T = \frac{|\bar{X} - \mu_0|}{\sigma/sqrtn}$ and let $S_1 = [c, \infty)$ for some $c > 0$. Then

$$\begin{aligned} P(T \in S_1|\mu) =& P\left(\frac{|\bar{X} - \mu_0|}{\sigma/\sqrt{n}} > c \Big| \mu\right) \\ =& P\left(\bar{X} \geq \mu_0 + \frac{\sigma}{\sqrt{n}}c \Big| \mu\right) + P\left(\bar{X} \leq \mu_0 - \frac{\sigma}{\sqrt{n}}c \Big| \mu\right) \\ =& P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{\mu_0 - \mu + \sigma/\sqrt{n}c}{\sigma/\sqrt{n}} \Big| \mu\right) + P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{\mu_0 - \mu - \sigma/\sqrt{n}c}{\sigma/\sqrt{n}} \Big| \mu\right) \\ =& 1 - \Phi\left(\frac{\mu_0 - \mu + \sigma/\sqrt{n}c}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{\mu_0 - \mu - \sigma/\sqrt{n}c}{\sigma/\sqrt{n}}\right) \end{aligned}$$

The following graph shows power curves for $\mu_0 = 0$, $n = 16$, and $\sigma^2 = 1$, for $c = 1, 2, 3$.



What is the probability of a type I error for each of the values of c above?

$$\pi(\mu = 0|\delta) = 1 - \Phi(c) + \Phi(-c) = 2\Phi(-c)$$

C	$\pi(\mu = 0\|\delta)$
1	0.317
2	0.046
3	0.003

Significance Level and Size

Def: Suppose there is a constant $\alpha_0 > 0$ so that $\pi(\theta|\delta) \leq \alpha_0$ for all $\theta \in \Omega_0$. Then δ is said to be a level α_0 test. The **size** of a test is

$$\alpha(\delta) = \sup_{\theta \in \Omega_0} \pi(\theta|\delta)$$

If H_0 is a simple hypothesis with $\Omega_0 = \theta_0$, then $\alpha(\delta) = \pi(\theta_0|\delta)$.

Cor: A test δ is a level α_0 test if and only if its size is at most α_0. If the null hypothesis for δ is simple, then δ is a level α_0 test if and only if $\pi(\theta_0|\delta) \leq \alpha_0$.

When choosing among a variety of hypothesis procedures, we might seek a test which has maximal power function for $\theta \in \Omega_1$ among all test that satisfy $\pi(\theta|\delta) \leq \alpha_0$ for $\theta \in \Omega_0$.

Ex 6: Returning to the coin flip, let's say that we flip $n = 100$ times, using the one-sided hypothesis $H_1 : \theta > 0.5$ (why?) and the critical region $S_1 = \{X \geq 55\}$. Then

$$\pi(\theta|\delta) = P(X \geq 55|\theta = .5) = 0.1587$$

If $\theta < 0.5$, then $\pi(\theta|\delta) < 0.1587$ and so the size of the test is 0.1587. Can we make δ a $\alpha = .05$ significance test?

Let's approximate $T(\mathbf{X})$ using the Normal distribution (valid by CLT) and consider $\delta : S_1 = \{T \geq c\}$.

$$\alpha(\delta) = \sup_{\theta \in \Omega_0} \pi(\theta|\delta) = \sup_{\theta \in \Omega_0} P(T \geq c|\theta) = \sup_{\theta \in \Omega} 1 - \Phi\left(\frac{c - 100\theta}{\sqrt{100\theta(1-\theta)}}\right)$$

Note that $1 - \Phi\left(\frac{c-100\theta}{\sqrt{100\theta(1-\theta)}}\right)$ is an increasing function of θ and a decreasing function of c.

Therefore,

$$\sup_{\theta \in \Omega} 1 - \Phi\left(\frac{c - 100\theta}{\sqrt{100\theta(1-\theta)}}\right) = 1 - \Phi\left(\frac{c - 50}{5}\right)$$

We now want to choose the smallest c so that

$$0.05 \geq \alpha(\delta) = 1 - \Phi\left(\frac{c-50}{5}\right)$$

Solving for c, we get

$$\frac{c-50}{5} \geq \Phi^{-1}(1-.05) \qquad c \geq 50 + 5\Phi^{-1}(1-.05) \approx 58.22$$

More generally, for significance α_0, if we let $T = \frac{X - n\theta_0}{\sqrt{\theta_0(1-\theta_0)/n}}$ be our test statistic, then we reject if

$$T \geq \Phi^{-1}(1-\alpha_0)$$

P-values

In a statistical investigation using a test procedure δ of level α, we will observe data $\mathbf{x}$ and report whether or not we rejected the null hypothesis. But in doing so, we fail to convey information about how **extreme** our data was. If we used a slightly modified test procedure at level $\alpha' < \alpha$, would we still have rejected the null hypothesis with this data? What is the most stringent level of evidence (i.e. the smallest value of α) at which we still would have rejected the H_0?

Def: (DeGroot and Schervish) The **p-value** is the smallest level α_0 such that we would reject the null-hypothesis at level α_0 with the observed data.

This definition is problematic. It is intended to allow us to define p-values for a very general class of test procedures, but it is actually too general, as the following counterexample shows:

Ex 7: Suppose we have a test procedure δ where the test statistic $T(\mathbf{X})$ has a continuous distribution. Suppose moreover that we observe the statistic $T(\mathbf{X}) = t$. Consider a different procedure δ' which has the same hypotheses, same partition of Ω, and same test statistic formula, but that uses rejection region $S_1 = \{t\}$. This procedure has size 0 and therefore, is a level $\alpha = 0$ test. However, it is also a test where we would reject the null hypothesis if we observed the statistic t. By the definition in Degroot and Schervish, the statistic t has a p-value of 0. But this is true for all t.

We need to focus on a reduced collection of procedures. This isn't a problem in practice, since we are usually only interested in deciding rejection regions from among similar options (for example, rejection regions for the form $[c, \infty)$ for some value of c).

Def: (STA 336) Let $\{\delta_c\}$ be a collection of hypothesis tests where δ_c has size α_c, such that for $\alpha_c < \alpha_{c'}$, if δ_c rejects H_0 when $\mathbf{x}$ is observed, then $\delta_{c'}$ also rejects H_0 when $\mathbf{x}$ is observed. The **p-value** for the observed sample $\mathbf{x}$ is the smallest level α_0 such that $\mathbf{x}$ is in the rejection region for δ_c.

Ex 8: Let $X_1, \ldots, X_{10}$ be a random sample from Unif$(\theta, \theta+1)$. To test $H_0 : \theta = 0$ versus $H_1 : \theta > 0$, we consider a collection of tests $\{\delta_c\}$ for $0 < c < 1$ where δ_c rejects H_0 if

$$\max\{X_1, \ldots, X_{10}\} \geq 1 \text{ or } \min\{X_1, \ldots, X_{10}\} \geq c$$

Suppose we observe $\mathbf{x}$ with $\min\{x_i\} = 0.1$ and $\max\{x_i\} = .5$. What is the p-value for this statistic?

Solution. Since H_0 is simple, the size of δ_c is $P(H_0 \text{ is rejected} | \theta = 0)$. Note that if $\theta = 0$, then $\max\{X_i\} \leq 1$. So

$$\alpha(\delta_c) = P(H_0 \text{ is rejected} | \theta = 0) = P(\min\{X_i\} \geq c | \theta = 0) = \prod P(X_i \geq c | \theta = 0) = (1-c)^{10}$$

Note that the size of the test $\alpha(\delta_c)$ decreases as c increases. Now, we reject H_0 for any procedure where $c \leq 0.1$. So the smallest sized test where we reject H_0 if we observed $\min\{x_i\} = 0.1$ is $\alpha = (1-.1)^n$. Hence, the p-value for $\mathbf{x}$ is $.9^{10} = 0.35$. ■

Informally, the p-value of a sample is the probability of obtaining another sample at least as extreme as the one observed, if the null hypothesis were true:

Thm: Consider a simple null hypothesis $H_0 : \theta = \theta_0$. Let $T(\mathbf{X})$ be a test statistic and consider a collection of tests $\{\delta_c\}$ of the form δ_c: "Reject H_0 if $T \geq c$". Suppose $T(\mathbf{x}) = t$. Then the p-value for this observed sample is

$$p\text{-value} = P(T \geq t|\theta_0).$$

More generally, for a composite null hypothesis $H_0 : \theta \in \Omega_0$, then the p-value for the observed sample is

$$p\text{-value} = \sup_{\theta \in \Omega_0} P(T \geq t|\theta).$$

Proof. Homework Exercise. □

Ex 9: Suppose we observe $X = 52$ in a sequence of 100 flips of a coin. We calculated before that

$$\alpha(\delta_t) = 1 - \Phi\left(\frac{t - 50}{5}\right)$$

and so the p-value of the statistic $X = 52$ is

$$p = 1 - \Phi\left(\frac{52 - 50}{5}\right) = 1 - \Phi\left(\frac{2}{5}\right) = 0.34$$

On the other hand, if we observed $X = 62$, then

$$p = 1 - \Phi\left(\frac{62 - 50}{5}\right) = 1 - \Phi\left(\frac{12}{5}\right) = 0.0082$$

Equivalence of Tests and Confidence Sets

Def: Let $\omega(\mathbf{X})$ be a random set, let g be a function, and let $0 \leq \gamma \leq 1$. Then $\omega(\mathbf{X})$ is said to be a γ-**confidence set** for $g(\theta)$ if, for every $\theta_0 \in \Omega$,

$$P(g(\theta_0) \in \omega(\mathbf{X}) \,|\, \theta = \theta_0) \geq \gamma$$

Thm: Let $\mathbf{X}$ be a random sample from a distribution with parameter θ, let $g(\theta)$ be a function.

1. Suppose that for each value θ_0 of θ, there is a level α_0 test δ_{θ_0} of the hypotheses

   $$H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0$$

   For each observed sample $\mathbf{X} = \mathbf{x}$, define

   $$\omega(\mathbf{x}) = \{\theta_0 \,|\, \delta_{\theta_0} \text{ does not reject } H_0 \text{ if } \mathbf{X} = \mathbf{x}\}$$

   Let $\gamma = 1 - \alpha_0$. Then the random set $\omega(\mathbf{X})$ is a γ-confidence set for $g(\theta)$.

2. Let $\omega(\mathbf{X})$ be a γ-confidence set for $g(\theta)$. For each value θ_0 of θ, construct a test δ_{θ_0} of the hypotheses

   $$H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0$$

   so that δ_{θ_0} does not reject H_0 iff $\theta_0 \in \omega(\mathbf{X})$. Then δ_{θ_0} is a $\alpha_0 = 1 - \gamma$ test of the hypotheses.

Ex 10: Suppose we wish to create a confidence interval for the probability p that coin flips Heads. We will consider a collection of tests of the hypotheses

$$H_0 : p = p_0 \qquad H_a : p \neq p_0$$

And will do so by flipping the coin n times and observing the number of heads X. For each value p_0, we can find a test δ_{p_0} of level α_0 the form "reject H_0 when $X \geq F^{-1}_{n,p_0}(1 - \alpha_0/2)$ or when $X \leq F^{-1}_{n,p_0}(\alpha_0/2)$

Suppose we observe data $X = x$ (i.e. x heads out of n). Among our tests, we do not reject H_0 if

$$F^{-1}_{n,p_0}(\alpha_0/2) < x < F^{-1}_{n,p_0}(1 - \alpha_0/2)$$

ALthough there is not a closed form for these p_0, for specific values of x, n, and α_0, we can search across values of p_0 for which this inequality is satisifed. These p_0 will then correspond to our confidence interval.

Ex 11: A batch of stout beer is best when it has an original gravity (OG) close to 1.071. Suppose the OG of beer is $N(\mu, \sigma^2)$. We sample 5 OG measurements from a batch of beer and find $\bar{x} = 1.0686$ and $s = 0.0064$. The γ-level confidence interval for μ is

$$1.068 \pm F_4^{-1}\left(\frac{1+\gamma}{2}\right)\frac{0.0064}{\sqrt{5}}$$

For each μ_0, we construct a $\alpha_0 = 1 - \gamma$ test of the hypotheses

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$$

where we reject H_0 if μ_0 is not in the preceding interval. Note moreover that this occurs exactly when

$$\frac{|1.086 - \mu_0|}{.0064/\sqrt{5}} \geq F_4^{-1}\left(\frac{1+\gamma}{2}\right)$$

Likelihood Ratio Tests

The likelihood ratio tests are probably the most frequently used tests in statistics. And there are several reasons for this:

1. In many cases, tests based on likelihood ratios have relatively high power (often, they are either the UMP or the UMP subject to extra restrictions)
2. The likelihood ratio is defined for general parametric models, so can be constructed without needing to find a specialized test for a given model
3. Test based on likelihood ratios are intuitive: we choose the hypothesis that is most consistent with the data!
4. Under general regularity conditions, the asymptotic distribution of the likelihood ratio is known.

Def: Suppose $\mathbf{X}$ is a sample from a distribution parameterized by θ, with joint PDF $f(\mathbf{x}|\theta)$. Consider general hypotheses of the form

$$H_0 : \theta \in \Omega_0 \quad H_1 : \theta \in \Omega_1$$

The **likelihood ratio** is the statistic

$$\Lambda(\mathbf{x}) = \frac{\sup_{\theta\in\Omega_0} f(\mathbf{x}|\theta)}{\sup_{\theta\in\Omega} f(\mathbf{x}|\theta)}.$$

Informally, the likelihood ratio test measures the relatively consistency of the data in the the null model compared to the full model. By construction, $0 \leq \Lambda(\mathbf{x}) \leq 1$. Values close to 0 indicate that the data was much less likely to occur under the null hypothesis than under the full model, while values close to 1 indicate either the data was as likely to occur under the null as the full model.

Hence, small values of $\Lambda(\mathbf{x})$ provide compelling evidence against the null hypothesis.

Def: A **Likelihood Ratio Test** of $H_0 : \theta \in \Omega_0$ against $H_1 : \theta \in \Omega_1$ is a procedure that rejects H_0 if $\Lambda(\mathbf{x}) \leq k$ for some constant k.

Ex 12: Suppose I want to test whether a given coin is fair after observing n flips. That is,

$$H_0 : \theta = 0.5 \quad H_1 : \theta \neq 0.5$$

Let X be the number of heads observed, with likelihood function

$$f(x|\theta) = \theta^x(1-\theta)^{n-x}$$

Note we dropped the coefficient that doesn't depend on θ. The likelihood ratio is then

$$\Lambda(x) = \frac{\theta_0^x(1-\theta_0)^x}{\sup_{\theta\in\Omega}\theta^x(1-\theta)^x} = \frac{0.5^n}{(x/n)^x(1-x/n)^{n-x}} = \frac{(0.5n)^n}{x^x(n-x)^{n-x}}$$

since the MLE for θ is $\frac{x}{n}$.

Suppose I want a test at the 0.05 level with $n = 100$. We can use R to compute the value of the statistic Λ for each value of x and use the PMF of x to find the PMF of Λ. Alternatively, we can use simulation to sample from x a large number of times to generate an approximate distribution for Λ, and compute the 0.05 quantile.

Doing the latter, I ended up with $k = 0.134$. Incidentally, this corresponds to getting more than 60 or less than 40 heads.