Lecture 6:

Hypothesis Testing – 2

The Permutation Test

The Permutation Test

- A nonparametric method for testing whether two distributions are the same. It is “exact” – not based on large sample approximations.

- Let $X_1, \ldots, X_m \sim F_X$ and $Y_1, \ldots, Y_n \sim F_Y$ be two independent samples. Hypotheses are: $H_0 : F_X = F_Y$ versus $H_1 : F_X \neq F_Y$ Let $T(x_1, \ldots, x_m, y_1, \ldots, y_n)$ be some test statistic, i.e:

$$T(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = |\overline{X}_m - \overline{Y}_n|$$

Let $N = m + n$ and consider forming all $N!$ permutations of the data $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. For each permutation, compute the test statistic T.

The Permutation Test

Let $N = m + n$ and consider forming all $N!$ permutations of the data $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. For each permutation, compute the test statistic T.

Denote these values $T_1, \ldots, T_{N!}$. Under H_0, each of these values is equally likely. The distribution that puts mass $1/N!$ at each T_j is called the **permutation distribution**. Let t_{obs} be the observed value of the test statistic. We reject when T is large, so the p-value is:

$$\text{p-value} = \frac{1}{N!} \sum_{j=1}^{N!} I(T_j > t_{\text{obs}})$$

The Permutation Test

- **Example**: The data are: $(X_1, X_2, Y_1) = (1{,}9{,}3)$. Let $T(X_1, X_2, Y_1) = |\overline{X} - \overline{Y}| = 2$. The permutations are:

permutation	value of T	probability
(1,9,3)	2	1/6
(9,1,3)	2	1/6
(1,3,9)	7	1/6
(3,1,9)	7	1/6
(3,9,1)	5	1/6
(9,3,1)	5	1/6

So the p-value is $\mathbb{P}(T > 2) = 4/6$.

The Permutation Test

- Of course, it is impractical to evaluate all $N!$ permutations for a large sample. We can approximate the p-value by randomly sampling from the set of permutations:

- So, the **Algorithm** for the permutation test is:

  1. Compute the observed value of the test statistic, t_{obs}

  2. Randomly permute the data. Compute the statistic again.

  3. Repeat the previous step B times. That gives $T_1, \ldots, T_B$

  4. Approximate the p-value with $\dfrac{1}{B}\displaystyle\sum_{j=1}^{B} I(T_j > t_{\text{obs}})$

The Likelihood Ratio test

The Likelihood Ratio Test

- The Wald test is useful for testing a scalar parameter. The likelihood ratio test is more general and can be used for testing a vector-valued parameter.

- Consider testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$. The **likelihood ratio statistic** is

$$\lambda = 2\log\left(\frac{\sup_{\theta\in\Theta}\mathcal{L}(\theta)}{\sup_{\theta\in\Theta_0}\mathcal{L}(\theta)}\right) = 2\log\left(\frac{\mathcal{L}(\widehat{\theta})}{\mathcal{L}(\widehat{\theta}_0)}\right)$$

  where $\widehat{\theta}$ is the MLE and $\widehat{\theta}_0$ is the MLE when θ is restricted to Θ_0

- There are other ways to define λ, but this is the most practical!

The Likelihood Ratio Test

- The LR test is most useful when Θ_0 consists of all θ values such that some coordinates of it are fixed at particular values:

- **Theorem**: Let $\theta = (\theta_1, \ldots, \theta_q, \theta_{q+1}, \ldots, \theta_r)$. Let $\Theta_0 = \left\{\theta : \ (\theta_{q+1}, \ldots, \theta_r) = (\tilde{\theta}_{q+1}, \ldots, \tilde{\theta}_r)\right\}$. Let λ be the LR test statistic. Under $H_0 : \ \theta \in \Theta_0$, we have $\lambda(x^n) \xrightarrow{d} \chi^2_{r-q,\alpha}$ where $r - q$ = (dimension of Θ)-(dimension of Θ_0). The p-value is $\mathbb{P}(\chi^2_{r-q} > \lambda)$.

- For example, $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$, and we want to test that $\theta_4 = \theta_5 = 0$. Then the limiting distribution has $5 - 3 = 2$ d.o.f.

The Likelihood Ratio Test

- **Example**: (Recall Mendel's peas). Mendel bred peas of 4 types. The number of each type is multinomial with $p = (p_1, p_2, p_3, p_4)$. His theory predicts that p equals $p_0 \equiv \frac{1}{16}(9,3,3,1)$. In $n = 556$ trials he observed $X = (315,101,108,32)$. For LR test, we have:

$$\lambda = 2\sum_{j=1}^{4} X_j \log \frac{\widehat{p}_j}{p_{0j}} = 2\left(315 \log\left(\frac{315/556}{9/16}\right) + \dots\right) = 0.48$$

Under H_1, there are 4 parameters. But they sum to 1, so the dim. of param. space = 3. Under H_0, there are no free params. So the limiting distribution of λ is χ^2_3, so p-value is $\mathbb{P}(\chi^2_3 > 0.48) = 0.92$

Multiple Testing

Multiple Testing

- Suppose we conduct several hypothesis tests, each at level α. For any one test, the chance of false rejection of the null is α. But the chance for **at least one** false rejection is much higher. This is the **multiple testing problem.** We will cover 2 methods to deal with this problem

- Consider m hypothesis tests: H_{0i} versus H_{1i} for $i = 1, \ldots, m$ and let $P_1, \ldots, P_m$ denote the m p-values for these tests.

- The **Bonferroni Method**: Given p-values $P_1, \ldots, P_m$, reject the null hypothesis H_{0i}, if $P_i < \dfrac{\alpha}{m}$

Multiple Testing

- **Theorem**: Using the Bonferroni method, the probability of falsely rejecting **any** null hypotheses is $\leq \alpha$.
  Proof idea: $\mathbb{P}(\bigcup_i R_i) \leq \sum_i \mathbb{P}(R_i)$

- The Bonferroni method is very conservative, trying to make it unlikely that you make even one false rejection. Sometimes it is more reasonable to control the **false discovery rate** (FDR) – the average fraction of false rejections among all rejections.

Multiple Testing

- Summarise everything in a table:

	H_0 Not Rejected	H_0 Rejected	Total
H_0 True	U	V	m_0
H_0 False	T	S	m_1
Total	$m - R$	R	m

- **Define** the **false discovery proportion** (FDP) as V/R, if $R > 0$, and 0 if $R = 0$. Then FDR $= \mathbb{E}(\text{FDP})$.

Multiple Testing

- The **Benjamini-Hochberg (BH) Method** is:

   1. Let $P_{(1)} < \ldots < P_{(m)}$ – ordered p-values.

   2. Define $\ell_i = \dfrac{i\alpha}{C_m m}$ where $C_m = \begin{cases} 1, & \text{if p-values - indep} \\ \sum_{i=1}^{m}(1/i) & \text{otherwise} \end{cases}$

      and $R = \max\left\{i: \ P_{(i)} < \ell_i\right\}$.

   3. Let $T = P_{(R)}$ – be the **BH rejection threshold**.

   4. Reject all null hypotheses H_{0i} for which $P_i \leq T$.

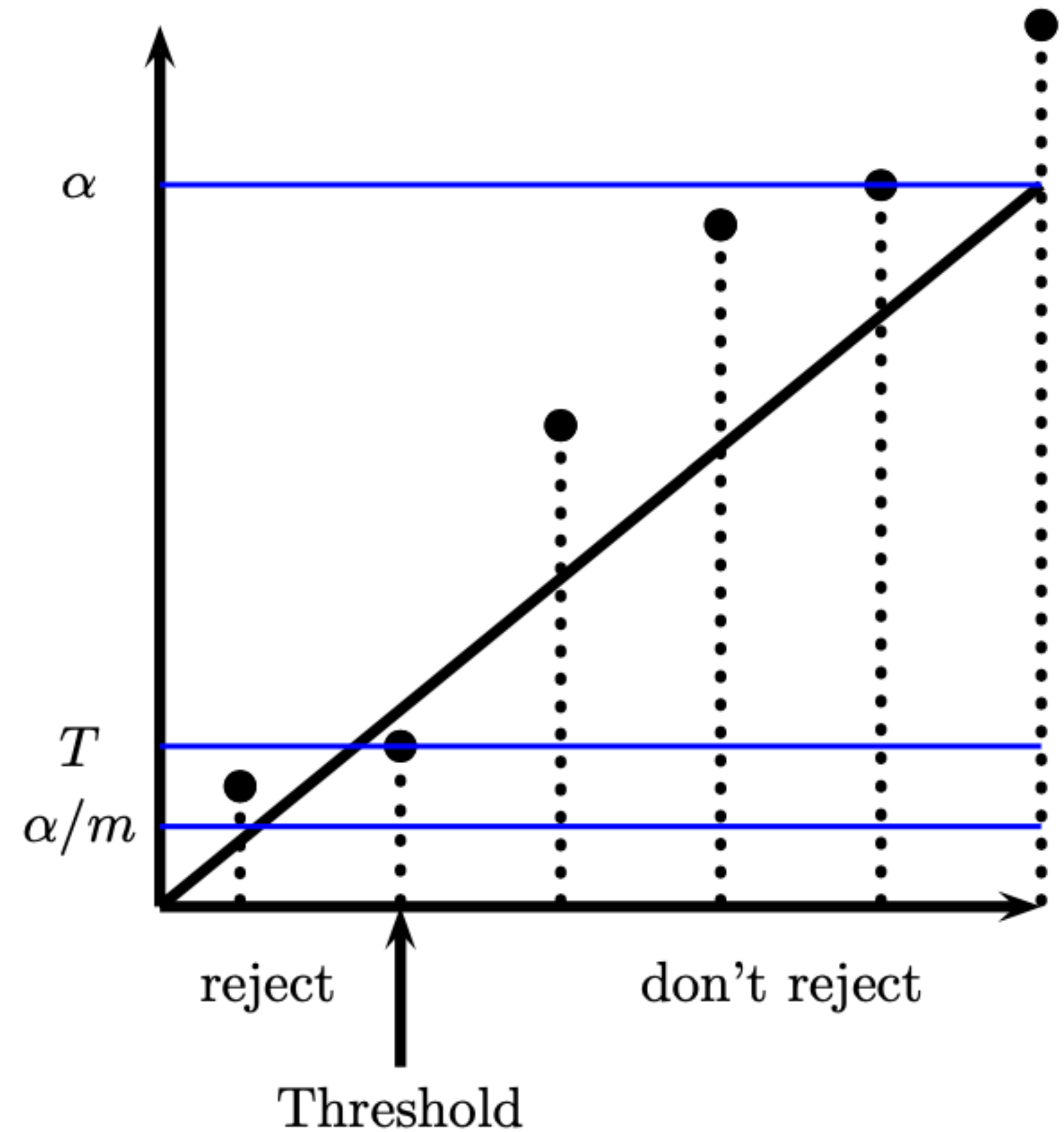
Multiple Testing

- **Theorem**: If the BH procedure is applied, then regardless of how many nulls are true, and regardless of the distribution of the p-values, when the null hypothesis is false,

$$\text{FDR} = \mathbb{E}(\text{FDP}) \leq \frac{m_0}{m}\alpha \leq \alpha$$

Multiple Testing

- **Example**: Suppose we have 6 tests, so 6 ordered p-values (vertical lines). 1) Without correcting for multiple testing, we would reject those with p-values $< \alpha$ – then 4 rejected. 2) Bonferroni rejects all whose p-values $< \alpha/m$ – then 0 rejected. 3) The BH threshold = last p-value under the line with slope α – then 2 rejected.



Goodness-of-fit tests

Goodness-of-fit tests

- Testing also arises when we want to check whether the data come from the assumed parametric model. There are many such tests, here is one.

- Let $\mathcal{F} = \left\{ f(x;\theta) : \ \theta \in \Theta \right\}$ be the parametric model. Suppose the data takes values on $\mathbb{R}$. Divide $\mathbb{R}$ into k disjoint intervals $I_1, \ldots, I_k$ and let $p_j(\theta) = \displaystyle\int_{I_j} f(x;\theta)\,dx$ – probability of falling in the interval I_j under assumed model. Let N_j observations fall into I_j. The likelihood of counts N_j is **multinomial**: $Q(\theta) = \displaystyle\prod_{j=1}^{k} p_j(\theta)^{N_j}$ Maximizing it yields estimates $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_s)$ of θ.

Goodness-of-fit tests

- The test statistic is $Q = \sum_{j=1}^{k} \frac{(N_j - np_j(\tilde{\theta}))^2}{np_j(\tilde{\theta})}$

- **Theorem**: Let H_0 be the null-hypothesis that the data are IID draws from our model $\mathscr{F} = \left\{ f(x;\theta) : \ \theta \in \Theta \right\}$. Under H_0, the statistic Q converges in distribution to χ^2_{k-1-s}. (This also gives an appropriate p-value).

- It is tempting to replace $\tilde{\theta}$ with the MLE, $\widehat{\theta}$. But this does not result in a statistic with χ^2_{k-1-s} limiting distribution. Although, some good things can be said in this case – a bound on the p-value due to Chernoff-Lehmann theorem, for example.

Goodness-of-fit tests

- Goodness-of-fit testing has limitations: if we reject H_0, we conclude we should not use the model; but if we do not reject H_0, we can not conclude that the model is correct – as always, we may have rejected because the test had low power. This is why it is generally better to use **nonparametric methods** when possible.

The Neyman-Pearson Lemma

The Neyman-Pearson Lemma

- In the special case of a simple test: $H_0 : \ \theta = \theta_0$ versus $H_1 : \ \theta = \theta_1$ we can say precisely what the most powerful test is
- **Theorem**: Let $T = \dfrac{\mathscr{L}(\theta_1)}{\mathscr{L}(\theta_0)} = \dfrac{\prod_{i=1}^n f(x_i; \theta_1)}{\prod_{i=1}^n f(x_i; \theta_0)}$. Suppose we reject H_0 when $T > k$. If we choose k so that $\mathbb{P}_{\theta_0}(T > k) = \alpha$, then this test is the most powerful size-α test.

The t-test

The t-test

- The t-test is due to Student's t-distribution:

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)\left(1+\frac{t^2}{k}\right)^{(k+1)/2}}$$ which, with d.o.f.

$k \to \infty$, tends to Normal, and with $k = 1$ reduces to Cauchy.

- If we want to test $H_0: \ \mu = \mu_0$ where $\mu = \mathbb{E}(X_i)$, we can use the Wald test. When the data is assumed **Normal** and **sample is small**, it is more common to use the t-test.

The t-test

- If we want to test $H_0: \ \mu = \mu_0$ where $\mu = \mathbb{E}(X_i)$, we can use the Wald test. When the data is assumed **Normal** and **sample is small**, it is more common to use the t-test.

- Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma)$, where μ, σ are both unknown. We want to test $\mu = \mu_0$ versus $\mu \neq \mu_0$. Let $T = \dfrac{\sqrt{n}(\overline{X}_n - \mu_0)}{S_n}$ where S_n^2 is the sample variance. For large samples, $T \approx \mathcal{N}(0,1)$ under H_0. But the exact distribution of T under H_0 is t_{n-1} (t-distribution witn n-1 degrees of freedom).