All of the material here is from the frequentist perspective.

1 Basics of Hypothesis Testing

Consider hypotheses of the form

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1.$$

- The rejection region R consists of values of the data $X = (X_1 \ldots, X_n)$ that are considered to be unlikely under H_0. Our decision rule for the hypothesis test is to reject H_0 if $X \in R$ or fail to reject/retain H_0 if $X \notin R$.
- Usually, the rejection region R is stated in terms of a test statistic $T(X)$ of the data, in which case

  $$R = \{x_1, \ldots, x_n : T(x_1, \ldots, x_n) > c\},$$

  where c is called a critical value.
- There are two types of errors that we can make when testing. A Type-I error happens when we reject the null even though it's true. A Type-II error happens when we fail to reject the null even though it's false. This can be summarized in the following table:

	Retain H_0 $X \notin R$	Reject H_0 $X \in R$
H_0 true	Correct	Type-I error
H_1 true	Type-II error	Correct

  The quality of a test is judged by its expected error rates:

  $$\text{Expected Type-I error rate} = \mathbb{P}(\text{Type-I error}) = \mathbb{P}_{H_0}(X \in R)$$
  $$\text{Expected Type-II error rate} = \mathbb{P}(\text{Type-II error}) = \mathbb{P}_{H_1}(X \in R).$$

- The power function with a rejection region R is defined as

  $$\beta(\theta) = \mathbb{P}_\theta(X \in R), \tag{1}$$

  i.e., the probability of rejecting the null for parameter value θ.
- The size of a test with rejection region R is defined as

  $$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(X \in R) = \sup_{\theta \in \Theta_0} \beta(\theta).$$

  A test is said to have level α if its size is less than or equal to α.

- Hypothetically, a good test has small $\beta(\theta)$ when $\theta \in \Theta_0$ and large $\beta(\theta)$ when $\theta \in \Theta_1$. But there are usually tradeoffs between the two. One way to handle this is the **Neyman-Pearson Paradigm**.

2 Neyman-Pearson Paradigm

The paradigm is as follows:

(a) Predefine a maximum Type-I error rate α that you are willing to tolerate. For a given $T(X)$, choose c such that the size of the test is less than or equal to α.

(b) Choose $T(X)$ so that the power of the test is as large as possible when $\theta \in \Theta_1$.

Under this paradigm, then the following theorem tells us what the best test is, when the hypotheses are simple.

Theorem 2.1. Neyman-Pearson Theorem. Suppose we test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$. Let

$$T(X) = \frac{L_n(\theta_1)}{L_n(\theta_0)} = \frac{f(X_1, \ldots, X_n; \theta_1)}{f(X_1, \ldots, X_n; \theta_0)},$$

the likelihood ratio. If we reject H_0 when $T > c$, with c chosen so that $\mathbb{P}_{\theta_0}(T > c) = \alpha$, then this is the most powerful size α test. That is, among all tests with size α, this test maximizes the power $\beta(\theta)$ for $\theta = \theta_1$.

In practice though, we are typically wanting to test a composite hypothesis (i.e., one that is not simple). How can we choose $T(X)$ in such a case? It can be challenging to find a most powerful test, and it may not even exist. So we will not cover that in this course and will instead cover a few commonly used testing paradigms.

3 A Few Common Testing Paradigms

3.1 (Generalized) Likelihood Ratio Test

This is a quite natural generalization of the other likelihood ratio test that was for simple hypotheses.

Definition 3.1. Likelihood Ratio Test. A likelihood ratio test (LRT) is a test with rejection region

$$R = \{x : \lambda(x) > c\},$$

where c is found so that the level of the test is α and

$$\lambda(X) = 2\log\left(\frac{\sup_{\theta\in\Theta} L_n(\theta)}{\sup_{\theta\in\Theta_0} L_n(\theta)}\right)$$

is the likelihood ratio.

Sometimes we can make exact probability calculations for λ, but sometimes it can be intractable. Instead, we can use the asymptotic distribution of λ, which is given by Wilks' Theorem.

Theorem 3.1. Wilks' Theorem. Under some assumptions (see lecture notes),

$$\lambda(X) \implies \chi^2_{k-q},$$

where k is the dimension of Θ and q is the dimension of Θ_0 (the dimension is the number of freely varying parameters).

Applying this, we will reject H_0 when λ is greater than the $(1-\alpha)$-th quantile of that chi-squared distribution.

3.2 Wald's Test

If we have an estimator that is asymptotically normal, then we can obtain an asymptotic α-level test based on that.

Definition 3.2. Wald Test. Consider testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. Let $\hat{\theta}_n$ be an estimator such that under H_0,

$$\frac{\hat{\theta}_n - \theta_0}{\hat{se}(\hat{\theta}_n)} \implies N(0,1).$$

The Wald Test rejects H_0 when $T(X) > z_{1-\alpha/2}$, where

$$T(X) = \left|\frac{\hat{\theta}_n - \theta_0}{\hat{se}(\hat{\theta}_n)}\right|.$$

If we want to use the MLE $\hat{\theta}_n^{MLE}$ as our estimator, then we know under some assumptions that $\hat{\theta}_n^{MLE} \to N(\theta, I_n(\theta)^{-1})$. So Wald's Test will be quite convenient to apply.

3.3 Permutation Test

The likelihood ratio test and Wald's Test are parametric. One example of a nonparametic test is a permutation test. In this course, we will focus on the case of comparing two groups, that is, if we have data $X_1, \ldots, X_{n_1}$ from distribution F_X and data $Y_1, \ldots, Y_{n_2}$ from a distribution F_Y.

Definition 3.3. Permutation Test. Consider the hypotheses $H_0 : F_X = F_Y$ vs. $H_1 : F_X \neq F_Y$. We reject when $T = g(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}) > c_\alpha$, where c_α is the $(1-\alpha)$-th quantile of the distribution under the null. The distribution under the null is found by considering all possible T_j orderings of the data.

4 P-Values

For any of the tests we've covered, they've given us a way to make a decision (reject or fail to reject H_0) using data. We can gain more information through p-values.

Definition 4.1. p-value. Suppose that for every $\alpha \in (0, 1)$, we have a size α test with rejection region R_α. Then the p-value is

$$\text{p-value}(X_{obs}) = \inf\{\alpha : X_{obs} \in R_\alpha\},$$

where X_{obs} is the observed data.

If we have a rejection region $R_\alpha = \{x : T(x) \geq c_\alpha\}$, then equivalently

$$\text{p-value}(X_{obs}) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X) \geq T(X_{obs})).$$

5 Duality of Hypothesis Tests & Confidence Intervals

There is a connection between hypothesis tests with point nulls and confidence intervals. Assume the hypotheses are of the form

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

Then:

- if we have an α level test, then a $1 - \alpha$ CI is the acceptance region for that test (the set of all possible values $\theta_0 \in \Theta$ for which we would fail to reject H_0).
- if we have a $1 - \alpha$ CI, then the acceptance region for an α-level test is the CI.

6 Multiple Testing

Please see the lecture notes for this topic.

Problem 1. Hypothesis Test for Repose Period Rate

The *repose period* for a volcanic eruption is the time between that eruption and the former eruption. Repose periods can range from less than a day, to hundreds of years, or even hundreds of thousands of years, depending on the type of volcano. It has been found that longer repose periods are associated with larger eruptions. Often the size of an eruption is quantified by the Volcanic Eruption Index (VEI), which ranges from 0 (non-explosive) to 8 (mega-colossal)
(`https://volcano.oregonstate.edu/faq/there-correlation-between-size-eruption-and-amount-time-volcano-last-erupted`).

One of Dr. George's main research specializations is volcanoes. He has heard several times that severe eruptions (i.e., those with VEIs of 3) have mean repose periods of about 10 years. He has a dataset of repose periods $X_1, \ldots, X_n$, which he will model as iid $\text{Exp}(\theta)$. He wants to test at $\alpha = 0.05$ the hypotheses

$$H_0 : \theta = \frac{1}{10}$$
$$H_1 : \theta \neq \frac{1}{10},$$

since $\mathbb{E}(X_i) = \frac{1}{\theta}$.

(a) Conduct a likelihood ratio test of these hypotheses, and calculate a confidence interval and p-value from that. Use the provided dataset (I simulated it hehe) in the section folder. Interpret the results.

(b) Repeat part (a) but for the Wald Test. Interpret the results.

(c) After this analysis, Dr. George becomes curious about the respose periods of the top 5 largest eruptions in the historical record, for Iceland vs. countries outside of Iceland. Suppose the values are
$\{100,000, 160,000, 350,000, 140,000, 210,000\}$ years for Iceland and
$\{90,000, 80,000, 150,000, 170,000, 110,000\}$ for the rest of the world. Code an exact permutation test in R at level 0.05 for whether these groups have the same distribution, based on the test statistic $T = |\text{median}(X_i) - \text{median}(Y_i)|$.

Solution

Part (a)

- The likelihood ratio is

$$\lambda(X) = 2 \log \left(\frac{\sup_{\theta \in \Theta} L_n(\theta)}{\sup_{\theta \in \Theta_0} L_n(\theta)} \right).$$

The full parameter space is $\Theta = (0, \infty)$, since $\theta > 0$ in the exponential distribution. The reduced parameter space under H_0 is simply $\Theta_0 = \{\frac{1}{10}\}$. The likelihood over the full parameter space is maximized at the MLE, which we have calculated in a previous section s $\hat{\theta} = \frac{1}{\bar{X}}$. The likelihood in the reduced parameter space is (trivially) maximized for $\frac{1}{10}$. We will also need the joint likelihood, which is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \theta e^{-\theta x_i} \text{ since all } X_i \text{ are iid} \\ &= \theta^n e^{-\theta \sum_{i=1}^{n} x_i}. \end{aligned}$$

We have then

$$\begin{aligned} \lambda(X) &= 2 \log \left(\frac{L(\hat{\theta})}{L(\frac{1}{10})} \right) \\ &= 2 \log \frac{\left(\frac{1}{\bar{X}^n}\right) e^{-\frac{\sum X_i}{\bar{X}}}}{\left(\frac{1}{10^n}\right) e^{-\frac{1}{10} \sum X_i}} \\ &= 2 \log \left(\frac{10^n}{\bar{X}^n} e^{-n} e^{\frac{1}{10} n \bar{X}} \right). \end{aligned}$$

Next, we must find the cutoff c that controls the Type-I error rate. There are two approaches for this.

Exact Approach: We can define the test statistic as $T(X) = \bar{X}$, since that is the only random quantity in λ. We will reject H_0 when $T(X) > c$, where c is chosen such that $\mathbb{P}_{\theta=1/10}(T > c) = \alpha$. For this problem, we could actually calculate this in closed form if we wanted to, using the fact that a sum of iid Exponential RVs is distributed as Gamma.

Asymptotic Approach: We can use Wilks' Theorem. For the degrees of freedom, we will need the dimensions of the parameter spaces. The dimension of Θ is 1, since θ is completely free to vary. The dimension of Θ_0 is simply 0, as the space consists only of a point. We have then that we will reject H_0 when $\lambda > d$, where d is the $1 - \alpha$-th quantile of the RV $\chi^2_{1-0} = \chi^2_1$.

- The p-value is

$$\begin{aligned} \text{p-value} &= \mathbb{P}_{\theta=1/10}(T(X) > T(X^{obs})) \\ &= \mathbb{P}(\chi^2_1 > T(X^{obs})). \end{aligned}$$

- By duality of CIs and hypothesis tests, a $(1-\alpha)$ CI for θ can be calculated as the acceptance region for this test.

- In R, λ^{obs} is calculated as about 7.095, while the critical value is $d \approx 3.841$. Since $\lambda^{obs} > d$, we reject H_0 at $\alpha = 0.05$. The p-value is about 0.0077, indicating very strong evidence

against the null hypothesis that $\theta = \frac{1}{10}$. A $(1-\alpha)100\%$ confidence interval corresponding to this test is about (0.059, 0.093). Dividing each endpoint by 1 gives (10.774, 16.949), which may be more interpretable. With 95% confidence then, we estimate that the mean repose period for severe eruptions is between about 10.774 and 16.949.

Part (b)

• A reasonable estimator to use for θ would be $\hat{\theta}_n = \frac{1}{\bar{X}}$, the MLE. We know that the asymptotic distribution of $\hat{\theta}_n$ is $N(\theta, I_n^{-1}(\theta))$, by a property of MLEs. We previously calculated the approximate variance as

$$\text{Var}(\hat{\theta}) = \frac{1}{I_n(\theta)} = \frac{\theta^2}{n}.$$

The Wald Test applies using this and has test statistic

$$\begin{aligned} T(X) &= \left| \frac{\hat{\theta}_n - \theta_0}{\hat{se}(\hat{\theta}_n)} \right| \\ &= \left| \frac{\frac{1}{\bar{X}} - \frac{1}{10}}{\frac{1}{\bar{X}\sqrt{n}}} \right|. \end{aligned}$$

We reject H_0 when $T > z_{1-\alpha/2}$.

• The p-value is

$$\begin{aligned} \text{p-value} &= \mathbb{P}_{\theta=1/10}(T(X) > T(X^{obs})) \\ &= \mathbb{P}(Z > T(X^{obs}) \\ &= 2\mathbb{P}(Z < -T(X^{obs})) \text{ (by symmetry of the normal dist. and since } T \text{ has an absolute value)} \\ &= 2\Phi(-T(X^{obs})). \end{aligned}$$

• A corresponding $(1-\alpha)100\%$ CI is simply

$$\begin{aligned} &\hat{\theta}_n \pm z_{1-\alpha/2}\hat{se}(\hat{\theta}_n) \\ &\frac{1}{\bar{X}} \pm z_{1-\alpha/2}\frac{1}{\bar{X}\sqrt{n}}. \end{aligned}$$

• In R, $T^{obs} \approx 2.944$, and the critical value is 1.96, so we reject H_0 at $\alpha = 0.05$. The p-value is about 0.0077, providing very strong evidence against the null hypothesis that $\theta = \frac{1}{10}$. A $(1-\alpha)100\%$ confidence interval corresponding to this test is about (0.058, 0.092). Dividing each endpoint by 1 gives (10.926, 17.3183). Therefore, with 95% confidence, we estimate that the mean repose period for severe eruptions is between about 10.926 and 17.318 years.

Overall, we can see that for this particular problem, the LRT and Wald Test led to the same conclusions.

Part (c)

See the R code, which has some comments detailing the steps too. The results are as follows. The observed test statistic is $T^{obs} = 50000$ and the critical value is $c = 60000$. Therefore, we fail to reject H_0 at level $\alpha = 0.05$. The p-value is about 0.191, which is larger than any standard level of significance. There is not enough evidence to reject the hypothesis that the biggest Icelandic eruptions have a different distribution of repose periods compared to the biggest eruptions in the rest of the world.

Problem 2. LRT for Regression Parameter

Suppose we are interested in understanding the association between housing prices and square footage. Let $Y_1, \ldots, Y_n$ denote the housing prices and $x_1, \ldots, x_n$ the corresponding square footages. We assume a simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ for each } i = 1, \ldots, n$$

where each x_i is considered fixed (non-random) and $\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} N(0, \sigma^2)$. All the parameters β_0, β_1, and σ^2 are unknown.

We randomly sample $n = 125$ houses for sale on Zillow in the U.S. (the dataset is in the section folder).

(a) We are interested in testing if $\beta_1 = 0$, because in that case there would be no association between house price and square footage under the model. The hypotheses then are

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0.$$

Conduct an LRT at level $\alpha = 0.1$.

(b) Now suppose instead that we were to test

$$H_0 : \beta_0 = \beta_1 = 0$$
$$H_1 : \text{not } H_0$$

with an LRT. Will λ and its asymptotic distribution be the same as in part (a)?

Solution

Part (a)

The parameter is $\theta = (\beta_0, \beta_1, \sigma^2)$, and the parameter spaces are

$$\Theta = \{(\beta_0, \beta_1, \sigma^2) : \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}, \sigma^2 \in [0, \infty)\}$$
$$\Theta_0 = \{(\beta_0, 0, \sigma^2) : \beta_0 \in \mathbb{R}, \sigma^2 \in [0, \infty)\}.$$

From Homework 3, the likelihood and log-likelihood were

$$L(\theta) = \nu^{-n/2} \exp\left(-\frac{1}{2\nu} \sum (Y_i - \beta_0 - \beta_1 x_i)^2\right)$$
$$l(\theta) = \frac{-n}{2} \log(\nu) - \frac{1}{2\nu} \sum (Y_i - \beta_0 - \beta_1 x_i)^2,$$

where $\nu = \sigma^2$. The MLE was found to be

$$\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2),$$

where

$$\begin{aligned}
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{x} \\
\hat{\beta}_1 &= \frac{s_{xY}^2}{s_x^2} \\
\hat{\sigma}^2 &= \frac{1}{n} \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.
\end{aligned}$$

So the LRT statistic is

$$\lambda(X) = 2 \log \frac{L(\hat{\theta})}{\sup_{\theta \in \Theta_0} L(\theta)}.$$

To get the denominator, we have to maximize the joint likelihood over σ^2 and β_0 when $\beta_1 = 0$ but the other parameters are free to vary. The log-likelihood when $\beta_1 = 0$ is

$$l = \frac{-n}{2} \log(\nu) - \frac{1}{2\nu} \sum (Y_i - \beta_0)^2.$$

Differentiating, we have

$$\begin{aligned}
\frac{\partial l}{\partial \beta_0} &= \frac{1}{\nu} \sum (Y_i - \beta_0) \\
\frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2\nu} + \frac{1}{2\nu^2} \sum (Y_i - \beta_0)^2.
\end{aligned}$$

Setting each equal to 0 and solving, we get

$$\begin{aligned}
\beta_0 &= \bar{Y} \\
\nu &= \frac{\sum (Y_i - \beta_0)^2}{n}.
\end{aligned}$$

Technically, would need to check the Hessian to confirm maximization.

Under H_0, we know $\lambda \approx \chi^2_{k-q}$, where $k = 3$ and $q = 2$ here. So we reject H_0 when λ exceeds the $(1 - \alpha)$-th quantile of this chi-squared distribution.

Putting all of this into R, we get that the observed LRT is $\lambda^{obs} = 19.21$ and $c = 3.84$, so we reject H_0 at level $\alpha = 0.1$. The p-value is calculated as 1.17×10^{-5}, indicating very high significance. This is not that surprising, since we know larger houses tend to cost more on average.

Part (b)

For this, the full parameter space Θ is the same as before. But the parameter space under the null will change to $\Theta_0 = \{(0, 0, \sigma^2) : \sigma^2 \in [0, \infty)\}$. The denominator of λ will involve maximizing the likelihood in this space. Additionally, if we use the chi-squared approximation, then the degrees of freedom for that will now be 3 - 1 = 2. So λ and its asymptotic distribution will be different from what we had in part (a).