Stat 231 Chapter 5: Hypothesis Testing

Statistics

David Duan, 2019 Fall

Contents

1 Hypothesis Testing

1.1 Null Hypotheses and Test Statistics

Statistical tests of hypotheses begin by specifying a single "default" hypothesis, called the *null hypothesis*, and then check whether the data collected is unlikely under this hypothesis. There is also an *alternative hypothesis*, which is the alternative to the null hypothesis.

Notation. The null and alternative hypothesis are denoted by H_0 and H_A (or H_1), respectively.

Ex. 1.1.1 Suppose we toss a coin 25 times to see if the coin is biased. Let Y be the number of heads in 25 trials. We assume Y has a binomial distribution with $n = 25$ and let θ be the probability of a head. If the coin is not biased, $\theta = 0.5$; if it is biased, then $\theta \neq 0.5$.

- The null hypothesis for this problem is "the coin is not biased", i.e., $H_0 : \theta = 0.5$.
- The alternative hypothesis for this problem is "the coin is biased", i.e., $H_A : \theta \neq 0.5$.

Intuitively, values $y \approx E[Y] = 12.5$ supports H_0 whereas y close to 0 or 25 provide evidence against H_0. To measure the strength of the evidence against H_0, we use a *test statistic*.

Def. 1.1.2 A *test statistic* or *discrepancy measure* is a function of the data $D = g(Y)$ which measures the degree of "agreement" between the data Y and the null hypothesis H_0.

Notation. Once we observed $Y = y$, the observed value of D is denoted $d = g(y)$.

Remark. D is a function of Y, so it is a random variable. We usually define D so that $d = 0$ represents the best possible agreement between the data and H_0, and large values of d indicate poor agreement.

1.2 p-Values

Computing $P(D = d)$ only tells us the probability of $D = d$; it does not tell us how much evidence we have for or against the H_0. Instead, we care about the probability of observing a value of D greater than or equal to the observed if H_0 were true.

Ex. 1.2.1 Let $D = |Y - 12.5|$. Suppose 10 tosses come up heads, so $d = 2.5$. The probability $P(D = 2.5)$ represents the probability of observing 10 heads or 15 heads, not the validity of H_0. Instead, we want to know what's the probability of $d = 2.5$ or larger happening given $H_0 : \theta = 0.5$ is true. Therefore, we want to know $P(D \geq 2.5; H_0) = P(|Y - 12.5| \geq 2.5)$ where $Y \sim Bin(25, 0.5)$. We can calculate this as

$$\begin{aligned} P(D \geq 2.5; H_0) &= P(Y \leq 10) + P(Y \geq 15) \\ &= 1 - P(11 \leq Y \leq 14) \\ &= 1 - \sum_{y=11}^{14} \binom{25}{y} (0.5)^{25} \\ &= 0.4244. \end{aligned}$$

How do we interpret this? Suppose we take a large number of coins *that are entirely fair* and toss each one 25 times. The result above tells us that *about* 42% *of those experiments would also observe a value of* $D = |Y - 12.5|$ *greater than equal to the* $d = 2.5$ *that we saw in our case!* This doesn't prove that our coin is not biased, but it does suggest that there is little evidence base on the observed data to support the alternative hypothesis.

The probability $P(D \geq 2.5) = 0.4244$ above is called the *p-value* of the test of hypothesis.

Def. 1.2.2 The p-value of the test hypothesis H_0 using test statistic D is $P(D \geq d; H_0)$.

In other words, the p-value is the probability of observing a value of the test statistic greater than or equal to the observed value of the test statistic assuming H_0 is true.

Remark.

1. We do not want $P(D = d; H_0)$ because we want to know the probability that, if H_0 were true, we'd see something *at least as extreme/unusual* as we actually observed.
2. A small p-value tells us that if H_0 is true, it would be unlikely to have observed data at least as surprising as the data we actually observed. Thus, small p-values provide evidence against H_0.

1.3 Statistical Tests of Hypotheses

Algorithm. 1.3.1 Here is a step-by-step guide through our hypothesis test:

1. Assume that H_0 will be tested using data Y.
2. Adopt a test statistic $D(Y)$, for which large values of D are less consistent with H_0.
3. Let $d = D(Y)$ be the corresponding observed value of D.
4. Calculate p-value: $P(D \geq d; H_0 \text{ is true})$.
5. Draw a conclusion based on the p-value.

Remark. You will learn how to pick an appropriate test statistic and carry out the p-value calculation in Section 2 and 3. To draw a conclusion based on the p-value, see Section 1.4.

1.4 Interpreting the p-Value

If d, the observed value of D, is large, and consequently the p-value $P(D \geq d; H_0)$ is small, then one of the following two statements is true:

1. H_0 is true but by chance we have observed an outcome that does not happen very often when H_0 is true.
2. H_0 is false.

The problem is, there is no specific threshold for p-values where if the p-value is smaller then we can conclude H_0 is false, and if the p-value is larger than we can conclude H_0 is true. p-values should only be interpreted in terms of their fundamental definition.

We do, nonetheless, provide interpretation guidelines to be used within the course.

Note. 1.4.1 (Interpreting p-Values) In general, values that are less than 0.01 are small which indicate that the observed data are providing strong evidence against H_0; values that are greater than 0.1 are large which indicate that we have not observed anything unusual when H_0 is true, so there is no evidence based on the observed data to suggest that H_0 is false.

$p - value$	Interpretation
$p - value > 0.10$	No evidence against H_0 based on the observed data.
$0.05 < p - value \leq 0.10$	Weak evidence against H_0 based on the observed data.
$0.01 < p - value \leq 0.05$	Evidence against H_0 based on the observed data.
$0.001 < p - value \leq 0.01$	Strong evidence against H_0 based on the observed data.
$p - value \leq 0.001$	Very strong evidence against H_0 based on the observed data.

Remark. We do NOT use terms like "reject" or "fail to reject" or say something is "statistically significant" in Stat 231. Instead, we focus on the interpretation of P-values in terms of the probabilities they represent.

Ex. 1.4.2 Consider an experiment with $n = 100$ coin tosses. Again, let $H_0 : \theta = 0.5$ and define the test statistic $D = |Y - 50|$. Suppose we observed $y = 40$ heads, so $d = |40 - 50| = 10$. By definition, the p-value is $P(D \geq 10)$ assuming H_0, i.e.,

$$P(D \geq 10; H_0) = P(|Y - 50| \geq 10) \text{ where } Y \sim Bin(100, 0.5).$$

Instead of computing $\sum_{i=1}^{100} P(Y = i)$, we could use the Gaussian approximation of binomial. Recall for $Y \sim Bin(n, \theta)$, if n is large, by CLT,

$$\frac{Y - n\theta}{\sqrt{n\theta(1-\theta)}} \sim G(0, 1) \quad \text{approximately.}$$

Thus,

$$\begin{aligned} P(D \geq 10; H_0) &= P(|Y - 50| \geq 10) \text{ where } Y \sim Bin(100, 0.5) \\ &= P\left(\frac{|Y - 50|}{\sqrt{100(0.5)(0.5)}} \geq \frac{10}{\sqrt{100(0.5)(0.5)}}\right) \\ &\approx P(|Z| \geq 2) \text{ where } Z \sim N(0, 1) \\ &= 2[1 - P(Z \leq 2)] = 0.04550. \end{aligned}$$

Using the chart above, we see there's evidence against H_0 based on the observed data.

Remark. To test other H_0, we just have to define a suitable discrepancy measure D, then compute $P(D \geq d)$ for our sample.

Note. 1.4.3 (p-Hacking) It is common in the scientific literature to see $p < 0.05$ used as the determinant of whether a hypothesis is or isn't rejected. However, this can lead to what's known as *p-hacking*, or *data dredging*: repeating experiments (or being selective with one's results) to falsely engineer a "significant" result. For example, I could perform 20 experiments, where in each experiment I tossed the coin 50 times. Even if H_0 is true, I'd still expect one of those 20 experiments to result in a p-value below 0.05. If I just present the 1 experiment where I get a significant p-value and don't mention the other 19, my result would be considered surprising. This is considered as p-hacking.

2 Hypothesis Testing for Gaussian Parameters

Suppose that $Y \sim G(\mu, \sigma)$ models a variate y in some population or process. A random sample $Y_1, Y_2, \ldots, Y_n$ is selected and we want to test hypotheses concerning one of the two parameters μ and σ. Recall the following estimators of μ and σ^2 from previous sections:

$$\tilde{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

2.1 Testing Mean Hypotheses with Unknown Standard Deviation

Recall the pivotal quantity

$$\frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

which was used to construct confidence intervals for μ. We will use this to construct a test of hypothesis for μ when σ is unknown.

Suppose we wish to test the hypothesis $H_0 : \mu = \mu_0$, where $\mu_0 \in \mathbb{R}$, against the alternative hypothesis that $\mu \neq \mu_0$. Values of $\overline{Y}$ which are either larger than μ_0 or smaller than μ_0 provide evidence against $H_0 : \mu = \mu_0$. Thus, we use the test statistic

$$D = \frac{|\overline{Y} - \mu_0|}{S/\sqrt{n}}.$$

We obtain the p-value using the fact that if $H_0 : \mu = \mu_0$ is true, then

$$\frac{|\overline{Y} - \mu_0|}{S/\sqrt{n}} \sim t_{n-1}.$$

Prop. 2.1.1 Let $d = \dfrac{|\bar{y} - \mu_0|}{s/\sqrt{n}}$ be the observed value of D in a sample with mean $\bar{y}$ and standard deviation s, then

$$\begin{aligned} p\text{-value} &= P(D \geq d; H_0) \\ &= P\left(\frac{|\overline{Y} - \mu_0|}{S/\sqrt{n}} \geq \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}}\right) \\ &= P(|T| \geq d) \quad T \sim t_{n-1} \\ &= 2[1 - P(T \leq d)]. \end{aligned}$$

Remark. Since values of $\overline{Y}$ which are larger or smaller than μ_0 provide evidence against the null hypothesis, this test is called a *two-sided test of hypothesis.* (See **Cor. 2.1.6** for *one-sided test of hypothesis.*)

Ex. 2.1.2 An inexpensive weight scale is tested by taking 10 weights of a known 1kg weight. Assume $Y_i \sim G(\mu, \sigma)$, $i = 1, 2, \ldots, 10$, where $Y_i = i$th measurement and μ represents the mean measurement in repeated weights of the 1kg weight using the scale. Let $H_0 : \mu = 1$. Suppose

$$\overline{y} = 1.0061, \quad \mu_0 = 1, \quad s = 0.029, \quad n = 10.$$

We can compute our test statistic:

$$d = \frac{|\overline{y} - \mu_0|}{s/\sqrt{n}} = \frac{|1.0061 - 1|}{0.0230/\sqrt{10}} = 0.839.$$

Our p-value is thus ($n = 10 \implies T \sim t_9$ below)

$$P(D \geq 0.839) = P(|T| \geq 0.839) = 2[1 - P(T < 0.839)] \approx 0.423.$$

Thus, there is no evidence against $H_0 : \mu = 1$ based on the observed data and we conclude there is no evidence the scale is under- or over-weighing.

Ex. 2.1.3 Suppose we repeat the experiment with another scale. Let $H_0 : \mu = 1$. Suppose

$$\overline{y} = 0.981, \quad \mu_0 = 1, \quad s = 0.0170, \quad n = 10.$$

Then

$$d = \frac{|\overline{y} - \mu_0|}{s/\sqrt{n}} = \frac{|0.981 - 1|}{0.0170/\sqrt{10}} = 3.534$$

and our p-value is ($n = 10 \implies T \sim t_9$ below)

$$P(D \geq 3.534) = P(|T| \geq 3.534) = 2[1 - P(T < 3.534)] \approx 0.00637.$$

There is evidence against $H_0 : \mu = 1$ and we conclude there is strong evidence the scale is under- or over-weighing.

Note. 2.1.4 (Statistical Significance vs. Practical Significance) Although there is strong evidence against H_0 for the second scale, the degree of bias in its measurement is not necessarily large enough be of practical concern. For example, a 95% confidence interval for the mean μ is given by $\overline{y} \pm 2.2622s/\sqrt{10} = 0.91 \pm 0.012 = [0.969, 0.993]$ (note: 2.2622 comes from the fact that $P(T \leq 2.2622) = 0.975$ given $T \sim t_9$). Thus, the bias in measuring the 1kg weight is fairly small (about 1% to 3%) and is not significant in practice.

If the evidence against H_0 is "statistically significant" (for some significance level, say 0.05), the size of the p-value does NOT imply how "wrong" H_0 is. The p-value just tells us how surprised we'd be by these data if the null hypothesis were true. A confidence interval, however, does indicate the magnitude and direction of the departure from H_0. If strong evidence against H_0 is found in a particular direction then this might suggest conducting further experiments to investigate this evidence.

To summarize, do not mix the following:

- *Statistical significance*: if our p-value is small (e.g., ≤ 0.05), then we have found evidence against H_0.
- *Practical significance*: look at the confidence interval and point estimate. If the size of the deviation from H_0 is small, we may not consider this of practical significance.

Note. 2.1.5 (Relationship between Hypothesis Testing and Interval Estimation)
Suppose we test $H_0 : \mu = \mu_0$ for $G(\mu, \sigma)$ data, then

$$\begin{aligned}
p\text{-value} \geq 0.05 &\iff P\left(\frac{|\overline{Y} - \mu_0|}{S/\sqrt{n}} \geq \frac{|\overline{y} - \mu_0|}{s/\sqrt{n}}\right) \geq 0.05 \\
&\iff P\left(|T| \geq \frac{|\overline{y} - \mu_0|}{s/\sqrt{n}}\right) \geq 0.05 && \text{where } T \sim t_{n-1} \\
&\iff P\left(|T| \geq \frac{|\overline{y} - \mu_0|}{s/\sqrt{n}}\right) \leq 0.95 && \text{where } T \sim t_{n-1} \\
&\iff \frac{|\overline{y} - \mu_0|}{s/\sqrt{n}} \leq a && \text{where } P(|T| \leq a) = 0.95 \\
&\iff \mu_0 \in [\overline{y} - as/\sqrt{n}, \overline{y} + as/\sqrt{n}]
\end{aligned}$$

which is a 95% confidence interval for μ! In other words, the p-value for testing $H_0 : \mu = \mu_0$ is greater than or equal to 0.05 if and only if the value $\mu = \mu_0$ is inside a 95% confidence interval for μ (assuming we use the same pivotal quantity).

More generally, suppose we use the same pivotal quantity to construct a confidence interval for a parameter θ and a test of the hypothesis $H_0 : \theta = \theta_0$. The parameter value $\theta = \theta_0$ is inside a $100q\%$ confidence interval for θ iff the p-value for testing $H_0 : \theta = \theta_0$ is greater than $1 - q$. (Or, if $p < 1 - q$, then a $100q\%$ confidence interval for μ will not contain μ_0 and vice versa.)

Cor. 2.1.6 (One-sided Test of Hypothesis for μ) Let $d = \dfrac{\overline{y} - \mu_0}{s/\sqrt{n}}$ be the observed value of D in a sample with mean $\bar{y}$ and standard deviation s, then

$$\begin{aligned}
p\text{-value} &= P(D \geq d; H_0) \\
&= P\left(\frac{\overline{Y} - \mu_0}{S/\sqrt{n}} \geq \frac{\overline{y} - \mu_0}{s/\sqrt{n}}\right) \\
&= P(T \geq d) \quad T \sim t_{n-1} \\
&= 1 - P(T \leq d).
\end{aligned}$$

2.2 Testing Variance Hypotheses with Unknown Mean

Recall the pivotal quantity

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

which was used to construct confidence intervals for σ. We will use this to construct tests of hypothesis for σ when μ is unknown.

Suppose we wish to test the hypothesis $H_0 : \sigma^2 = \sigma_0^2$, where $\sigma_0 \in \mathbb{R}$, against the alternative hypothesis that $\sigma^2 \neq \sigma_0^2$. If H_0 is true, then

$$U = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

Remark. Previously, we have defined our discrepancy measure so that larger values indicate evidence against H_0, while smaller values would suggest the data are consistent with H_0. This is because our test statistic is an absolute value

$$D = \frac{|Y - \mu_0|}{S/\sqrt{n}}$$

and for an observed value d, we compute p-value as $P(D \geq d)$. We could instead have defined

$$D' = \frac{Y - \mu_0}{S/\sqrt{n}}$$

and for an observed value d, compute p-value as $P(D' \leq -|d'|) + P(D' \geq |d'|)$. The absolute value formula works because the t-distribution is symmetric, so $P(D' \leq -|d'|) = P(D' \geq |d'|)$.

The chi-squared distribution is not symmetric about its mean, which makes the determination of "large" and "small" values slightly more challenging.

Prop. 2.2.1 Let $u = \dfrac{(n-1)s^2}{\sigma_0^2}$ be the observed value of U in a sample with unknown mean and standard deviation s,

- If $P(U \leq u) < 0.5$, then p-value $= 2P(U \leq u)$.
- If $P(U \geq u) < 0.5$, then p-value $= 2P(U \geq u)$.

where $U \sim \chi^2_{n-1}$.

Ex. 2.2.2 Suppose we wish to test the hypothesis $H_0 : \sigma^2 = 0.26$ given sample size $n = 30$ and sample variance $s^2 = 0.311$. Then the test statistic is

$$U = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{29}.$$

Plug in n and s, we get

$$u = \frac{29 \cdot 0.311}{0.26} \approx 34.69.$$

Given $U \sim \chi^2_{29}$, $P(U \leq u) = P(U \leq 34.69) \approx 0.785$; $P(U \geq u) = 1 - P(U \leq 34.69) = 0.215$. Since the latter is smaller than 0.5, the p-value for this case is $2 \cdot 0.215 = 0.43$.

3 Likelihood Ratio Test Statistic

When a pivotal quantity exists, it is usually straightforward to construct a test of hypothesis as we have seen in Section 2. When a pivotal quantity does not exist, then a general method for finding a test statistic with good properties can be based on the likelihood function.

3.1 Motivation

Recall from Chapter 2, we used likelihood functions to gauge the plausibility of parameter values in the light of the observed data. We could based a test of hypothesis on a likelihood value or, in comparing the probability of two values, a ratio of the likelihood values.

Suppose there are two estimates θ_0 and θ_1 for θ. Having some data y at hand, we could take a look at the ratio

$$\frac{L(\theta_0)}{L(\theta_1)}.$$

If this ratio is much greater than one, then the data support the value θ_0 more than θ_1.

Now suppose we want to test the plausibility of hypothesized value θ_0 against an unspecified alternative, which is usually the MLE $\hat{\theta}$, so we replace θ_1 with $\hat{\theta}$. Then the resulting ratio is just the value of the relative likelihood function at θ_0:

$$R(\theta_0) = \frac{L(\theta_0)}{L(\hat{\theta})}.$$

If $R(\theta_0)$ is close to one, then θ_0 is plausible given the observed data, but if $R(\theta_0)$ is very small and close to zero, then θ_0 is not very plausible and this suggests evidence against H_0.

Therefore, the corresponding random variable

$$\frac{L(\theta_0)}{L(\tilde{\theta})}$$

appears to be a natural statistic for testing $H_0 : \theta = \theta_0$.

3.2 Likelihood Ratio Test Statistic

To determine p-values, we need the sampling distribution of $L(\theta_0)/L(\tilde{\theta})$ under H_0. Recall the likelihood ratio statistic from Chapter 4:

$$\Lambda(\theta_0) = -2\log\left[\frac{L(\theta_0)}{L(\tilde{\theta})}\right]$$

which is a one-to-one function of $L(\theta_0)/L(\tilde{\theta})$. If $H_0 : \theta = \theta_0$ is true, then $\Lambda(\theta_0)$ has approximately χ^2_1 distribution. Note that small values or $R(\theta_0)$ correspond to large observed values of $\Lambda(\theta_0)$ (because of the negation) and therefore large observed value of $\Lambda(\theta_0)$ indicates evidence against the hypothesis $H_0 : \theta = \theta_0$.

Prop. 3.2.1 To determine the p-value,

1. Calculate the observed value of $\Lambda(\theta_0)$, denoted $\lambda(\theta_0)$:

$$\lambda(\theta_0) = -2\log\left[\frac{L(\theta_0)}{L(\hat{\theta})}\right] = -2\log R(\theta_0)$$

   where $R(\theta_0)$ is the relative likelihood function evaluated at $\theta = \theta_0$.

2. We can then approximate the p-value as

$$\begin{aligned} p\text{-value} &\approx P[W \geq \lambda(\theta_0)] & W \sim \chi^2_1 \\ &= P(|Z| \geq \sqrt{\lambda(\theta_0)}) & Z \sim G(0,1) \\ &= 2[1 - P(Z \leq \sqrt{\lambda(\theta_0)})] \end{aligned}$$

Ex. 3.2.2 (Likelihood Ratio Test Statistic for Binomial Model) Let's first derive the formula for $\lambda(\theta_0)$. Recall the relative likelihood function for the binomial is

$$\begin{aligned} R(\theta) &= \frac{L(\theta)}{L(\hat{\theta})} \\ &= \frac{\theta^y(1-\theta)^{n-y}}{\hat{\theta}^y(1-\hat{\theta})^{n-y}} \\ &= \left(\frac{\theta}{\hat{\theta}}\right)^y \left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n-y} \qquad 0 \leq \theta \leq 1 \end{aligned}$$

The likelihood ratio test statistic for testing $H_0 : \theta = \theta_0$ is

$$\Lambda(\theta_0) = -2\log\left[\frac{L(\theta_0)}{L(\tilde{\theta})}\right] = -2\log\left[\left(\frac{\theta}{\tilde{\theta}}\right)^y \left(\frac{1-\theta}{1-\tilde{\theta}}\right)^{n-y}\right]$$

where $\tilde{\theta} = Y/n$ is the maximum likelihood estimator of θ. The observed value of $\Lambda(\theta_0)$ is

$$\lambda(\theta_0) = -2\log R(\theta_0) = -2\log\left[\left(\frac{\theta_0}{\hat{\theta}}\right)^y \left(\frac{1-\theta_0}{1-\hat{\theta}}\right)^{n-y}\right]$$

where $\hat{\theta} = y/n$. If $\hat{\theta}$ is close to θ_0, then $R(\theta_0)$ will be close to 1 and $\lambda(\theta_0)$ will be close to 0.

Suppose we are given $n = 200$, $y = 110$, and $\hat{\theta} = 0.55$. Then the likelihood ratio statistic for testing $H_0 : \theta = 0.5$ is

$$\lambda(0.5) = -2\log R(0.5) = -2\log(0.367) = 2.003.$$

Note that $R(0.5) = 0.367 > 0.1$ tells us that $\theta = 0.5$ is a plausible value of θ.

The approximate p-value for testing $H_0 : \theta = 0.5$ is

$$\begin{aligned} p\text{-}value &\approx P(W \geq 2.003) & W \sim \chi^2_1 \\ &= 2[1 - P(Z \leq \sqrt{2.003})] & Z \sim G(0,1) \\ &= 2[1 - P(Z \leq 1.42)] \\ &= 2(1 - 0.9222) = 0.1556. \end{aligned}$$

Based on our guideline above, there is no evidence against $H_0 : \theta = 0.5$ based on the data.

Remark. More examples can be found in course notes.