Lecture 7: Hypothesis Testing

Defn (Hypothesis):

A statement regarding the value of a parameter, θ.

Null hypothesis: $H_0: \theta = \theta_0$, specified.

Alternative hypothesis: $H_a: \theta < \theta_0$ or
$\theta > \theta_0$ or
$\theta \neq \theta_0$

Decision rule: rule for deciding ~~on~~ when to reject H_0 or fail to reject H_0.

Significance level: amount of "evidence" needed to reject H_0.

Example 1: Judicial analogy

In a criminal count, you put defendants on trial because you suspect they are guilty of a crime. But how does the trial proceed?

Determine the null and alternative hypotheses

H_0: defendant is not guilty

H_a: defendant is guilty

Select a significance level as the amount of evidence needed to convict.

In a court of law, the evidence must prove guilt "beyond a reasonable doubt".

Collect evidence and then use a decision rule to make a judgement. If the evidence is (a) sufficiently strong, reject H_0

Note: Failing to prove guilt does not prove that the defendant is innocent.

Example 2: Coin Analogy

Suppose you want to know whether a coin is fair. You cannot flip it forever, so you decide to take a sample. Flip it five (5) times and count the number of heads and tails.

H_0: coin is fair vs H_a: coin is not fair

Significance level: if you observe 5 heads in a row or five tails in a row, you conclude the coin is not fair; otherwise, you decide there is not enough evidence to show the coin is not fair.

You flip the coin five times and count the number of heads and tails.

You evaluate the data using your decision rule and make a decision that there is (a) enough evidence to reject H_0 or (b) not enough evidence to reject H_0

Now, in mathematical language:

$H_0: p = 1/2$ versus $H_a: p \neq 1/2$.

Let $X = \#$ heads in $n = 5$ tosses.

Then $X \sim b(n=5, p=1/2)$ under H_0.

Reject H_0 if $x = 5$

Types of Errors:

	Actual	
Decision	H_0 True	H_0 False
Reject H_0	Type I	Correct
Fail to Reject H_0	Correct	Type II

$P(\text{Type I error}) = \alpha$

$P(\text{Type II error}) = \beta$

Legal example: $\alpha =$ probability of concluding defendant is guilty when they are innocent

$\beta =$ probability of failing to find the person guilty when they are guilty

Coin example; $\alpha = ?$

$\beta = ?$

Power of a statistical test $= 1 - \beta$ i.e

the probability that you correctly reject H_0.

Let us consider the modified coin expt.
Flip a fair coin 100 times and decide whether it is fair:

(1) 55 heads/45 tails $\Rightarrow$ difference = 10

(2) 40 heads/60 tails $\Rightarrow$ difference = 20

(3) 37 heads/63 tails $\Rightarrow$ difference = 26

(4) 15 heads/85 tails $\Rightarrow$ difference = 70.

If you flip a coin 100 times and count the number of heads, you do not doubt that the coin is fair if you observe exactly 50 heads. However, you might be

a. Somewhat skeptical that the coin is fair if you observe 40 or 60 heads

b. even more skeptical that the coin is fair if you observe 37 or 63 heads

c. highly skeptical that the coin is fair if you observe 15 or 85 heads

Here, the greater the difference btwn the number of heads and tails, the more evidence you have that the coin is not fair.

Defn (p-value):

A p-value is the probability of observing a value as extreme or more extreme than the one observed.

Example: Ho: coin is fair and you observe 40 heads (60 tails), then

the p-value is the probability of observing a difference in the number of heads and tails of 20 or more from a fair coin tossed 100 times = 0.06

For 55 heads/45 tails, p-value $= 0.37$

For 37 heads/63 tails, p-value $= 0.01$

Lastly, for 15 heads/85 tails, p-value < 0.001

If the p-value is <u>large</u>, you would see often a difference this large in experiments with a fair coin. But if the p-value is <u>small</u>, you would rarely see differences this large from a fair coin, giving the evidence that the coin is not fair.

<u>Note:</u> $P(X < 40) + P(40 \le X \le 60) + P(X > 60) = 1$

where $X \sim \text{binomial}(p = 1/2, n = 100)$

Using the normal approximation to the binomial,

$$P(40 \le X \le 60) = P(-2 \le Z \le 2)$$
$$= 2P(Z < 2) - 1$$
$$= 0.9544$$

But $P(X < 40) = P(X > 60)$ by symmetry

$\Rightarrow 2P(X > 60) + 0.9544 = 1 \Rightarrow P(X > 60) = 0.0228$

Required p-value $= 0.0456 + P(X = 60) = 0.0456 + 0.0108 = 0.0564 \approx 0.06$

In statistical hypothesis testing, the significance level $= \alpha$ (Type I error rate) and the strength of the evidence is measured by a p-value.

Decision rule: Reject H_0 if p-value $< \alpha$
Fail to reject H_0 if p-value $\geq \alpha$

Some Common Statistical Tests

1. $H_0: \mu = \mu_0$ vs $H_a: \mu \neq \mu_0$ $\quad X \sim N(\mu, \sigma^2)$

Test statistic: $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$

P-value $= P(T > |t|) = 2P(T > t)$

Reject H_0 if p-value $< \alpha$

Example 1:

It is suspected that a machine, used for drilling plastic bottles with a net volume of 16.0 ounces, on average, does not perform according to specifications. An engineer will collect 15 measurements and will reset the machine if there is evidence that the mean fill volume is different from 16 ounces. The resulting data yields $\bar{x} = 16.0367$ ounces and $s = 0.0551$ ounces. Test the hypothesis

$$H_0: \mu = 16 \quad vs \quad H_a: \mu \neq 16 \quad at \ \alpha = 0.05$$

Solution:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{16.0367 - 16}{0.0551/\sqrt{15}}$$

$$= 2.58$$

$$p\text{-value} = P(T > 2.58) \quad < 0.025$$

$\Rightarrow$ we reject H_0.

2. $H_0: \mu_1 = \mu_2$ vs $H_a: \mu_1 \neq \mu_2$

where $X_1 \sim N(\mu_1, \sigma^2)$ and

$X_2 \sim N(\mu_2, \sigma^2)$

Test statistic: $T = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t(n_1+n_2-2)$

Reject Ho if p-value $= P(T > |t|)$

Here $s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-2)s_2^2}{n_1+n_2-2}$

where $\bar{x}_i$, s_i^2 and n_i are the sample mean, sample variance and sample size for each sample i, $i = 1, 2$.

Example 2: $\bar{x}_1 = 24.2$, $s_1 = \sqrt{10}$, $n_1 = 35$

$\bar{x}_2 = 23.9$, $s_2 = \sqrt{14.3}$, $n_2 = 40$

Then $t = \dfrac{24.2 - 23.9}{\sqrt{12.3\left(\frac{1}{35} + \frac{1}{40}\right)}} = 0.369$

p-value $= P(T > 0.369) = 0.1785 > 0.025$

Neyman-Pearson Lemma

Any hypothesis testing problem involves a trade-off between Type I and Type II error probabilities.

Question: Since many α-level tests could be possible in any hypothesis testing problem, how do we know whether or not we are using the best possible test?

Answer: Find the MOST POWERFUL test of level α

= Likelihood ratio test

Neyman-Pearson Lemma:

For a fixed k ($0 \leq k < \infty$), consider a test that rejects $H_0: \theta = \theta_0$ vs $H_a: \theta = \theta_1$ when

$$\frac{L(\theta_1 \mid x_1, x_2, \ldots, x_n)}{L(\theta_0 \mid x_1, x_2, \ldots, x_n)} > k$$

Let $\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 \mid \theta_0)$

Then this test is the most-powerful test of size α i.e it maximizes

$$\text{Power} = P(\text{Reject } H_0 \mid \theta_1)$$

among all tests with same level α.

<u>Example</u>:

Derive the most powerful α-level test of $H_0: \mu = \mu_0$ vs $H_a: \mu = \mu_1$ where $X \sim N(\mu, \sigma^2)$, (σ^2 known).

<u>Solution</u>:

$$\frac{L(\mu_1 \mid x_1, x_2, \ldots, x_n)}{L(\mu_0 \mid x_1, x_2, \ldots, x_n)} = \frac{\exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\mu_1)^2\right\}}{\exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\mu_0)^2\right\}}$$

$$= \exp\left\{\frac{n}{2\sigma^2}\left[(\bar{x}-\mu_0)^2 - (\bar{x}-\mu_0)^2 - (\bar{x}-\mu_1)^2\right]\right\}$$

$$= \exp\left\{\frac{n}{2\sigma^2}\left[2\bar{x} - (\mu_0 + \mu_1)\right](\mu_1 - \mu_0)\right\}$$

$$\Rightarrow \bar{x} > \frac{\sigma^2 \ln k}{n(\mu_1 - \mu_0)} + \frac{\mu_0 + \mu_1}{2} = k^*$$

We choose k^* so that

$$P(\bar{X} > k^* \mid \mu = \mu_0) = \alpha.$$

$$\Rightarrow \quad k^* = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.$$