VIII. TESTS OF SIGNIFICANCE

VIII.20 Tests of Significance. One-Sample z-Test and t-Test.

- The idea of "hypothesis testing"

  Suppose we are interested in a target parameter \( \theta \) associated with the probability distribution of a chance variable \( X \) measured in some population. For example, the mean \( \mu \).

  - And suppose we acquire relevant data:

    \[ X_1, X_2, \ldots, X_n \]

  - One possible approach (that we have seen) is to develop from the data a sample estimator \( \hat{\theta} \) of \( \theta \). For example, \( \bar{X} \) for estimation of \( \mu \).

  - On the other hand, perhaps we already have a special value of \( \theta \) in mind, say \( \theta_0 \). This might derive from scientific considerations, for example. Then we would like to investigate how strongly the data supports the "hypothesis" that \( \theta = \theta_0 \).

- Example Is the coin you have a "fair" coin? Is the probability of Heads \( p = \frac{1}{2} \)? Target parameter: \( p \).

  Hypothesized value: \( p_0 = \frac{1}{2} \).

  Toss the coin \( n = 50 \) times, get \( k = 29 \) Heads.

  Then \( \hat{p} = \frac{29}{50} = 0.58 \). We ask:

  - Is our data evidence for or against \( p = \frac{1}{2} \)?
  - How strongly?
The structure of "hypothesis testing"

1. We state a **null hypothesis**:
   
   \[ H_0 : \theta = \theta_0 \text{ (specified value)} \]

2. We state an **alternative hypothesis** of special interest. When \( \theta \) is a numerical value, the alternative hypothesis can be one of 3 possibilities:
   
   "\( H_A : \theta \neq \theta_0 \)" (2-sided)
   
   or "\( H_A : \theta < \theta_0 \)" (one-sided)
   
   or "\( H_A : \theta > \theta_0 \)" (one-sided)

3. We obtain data and use a **test statistic** which suggests whether or not the data supports the chosen alternative \( H_A \) more than the null hypothesis \( H_0 \). On this basis, we either
   
   (a) Accept \( H_0 \) as true
   
   or
   
   (b) Reject \( H_0 \) in favor of the chosen \( H_A \).

**Note.** If \( H_A \) is one-sided, we reject \( H_0 \) only if the data favors that \( H_A \) (not if it favors the other one-sided \( H_A \)).

The test statistic is usually designed so that if \( H_0 \) is true, then the test statistic takes values not too far from 0. If the test statistic takes a value rather far away from 0 and consistently with the alternative \( H_A \), then we reject \( H_0 \) in favor of \( H_A \). Otherwise we accept \( H_0 \).
Examples of "hypothesis testing"

(i) Jurisprudence

H₀ : "The accused is innocent"
Hₐ : "The accused is guilty."

In our judicial system, the accused is assumed innocent unless the evidence against this assumption is sufficiently strong "beyond a reasonable doubt." This is decided by a jury of "peers." This system favors letting a guilty person go free over imprisoning an innocent person.

(ii) Is the coin "fair"?

H₀ : p = 1/2
Hₐ : p ≠ 1/2 (two-sided alternative)

(iii) Is the average years of schooling in Town A the same as the national average?

H₀ : μ = μ₀ (the national average)
Hₐ : μ < μ₀ (one-sided alternative)

This structure corresponds to the following thinking: If μ equals or exceeds the national average, then no action is desired. If, however, μ is less than the national average, some action might be taken to strengthen the educational resources in this town.

Hence, in effect, we are testing
H₀ : no action needed
versus
Hₐ : special action needed.
Examples of test statistics

(i) Is the coin fair?

\[ H_0: \pi = \frac{1}{2} \quad \text{versus} \quad H_a: \pi \neq \frac{1}{2} \]

Data: \( k = 29 \) Heads in \( n = 50 \) tosses

\[ \hat{\pi} = \frac{29}{50} = 0.58 \]

Test statistic

\[ Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \]

If \( H_0 \) true, we expect a small difference.
If not, then a big difference.

\[ \Rightarrow \]

\[ Z = \frac{(\text{estimated value of } \pi) - (\text{hypothesized value of } \pi)}{(SE(\hat{\pi}) \text{ if } H_0 \text{ true})} \]

\[ = \frac{0.58 - 0.5}{\sqrt{0.5 \times 0.5 \times \frac{50}{1}}} \]

\[ = \frac{0.08}{0.0907} \]

\[ = 1.13 \]

The above illustrates design of a suitable test statistic. Later we will see how to interpret the observed value of 1.13 as evidence for or against \( H_0 \).
(ii) Average years of schooling ($\mu$) in Town A

$H_0: \mu = 12.3 \quad \leftarrow \text{The national average, } M_0$

$H_A: \mu < 12.3 \quad \leftarrow \text{We are concerned with whether } \mu < M_0 \text{ or not.}$

**Data:** $\bar{X} = 11.6 \text{ yrs} \quad \text{SD} = 4.1 \text{ yrs}, \text{ based on sample size 400}$

**Test statistic**

$$\frac{\bar{X} - M_0}{\sqrt{\frac{\text{SD}^2}{n}}}$$

$$= \frac{11.6 - 12.3}{\sqrt{\frac{(4.1)^2}{400}}}$$

$$= -0.7$$

$$= -0.205$$

$$= -3.41$$

The above illustrates design of a suitable test statistic. Later we will interpret the value $-3.41$ as evidence for or against $H_0$. 
Measuring the evidence in the data: the "significance level" or "probability value" or "p-value" or "P-value"

The P-value is the chance of getting a more extreme value of the test statistic than the value actually obtained.

This chance is computed using the probability model for the test statistic if $H_0$ is true.

"More extreme" means consistent with $H_A$. If one-sided in a certain direction, then more extreme in that direction. If two-sided, then more extreme in magnitude.

Example: $H_0: \mu = \mu_0$ versus $H_A: \mu < \mu_0$

**This area is the P-value**

**Density curve for test statistic if $H_0$ true**

**Value obtained for test statistic**

Smaller P-value corresponds to stronger evidence against $H_0$. 
Illustration of P-value in Examples

(c) Is the coin fair?

\[ H_0: p = \frac{1}{2} \text{ versus } H_A: p \neq \frac{1}{2} \]

Test statistic:

\[ \frac{\hat{p} - \bar{p}_0}{\sqrt{\frac{\bar{p}_0(1-\bar{p}_0)}{n}}} = \cdots = 1.13 \quad \text{(recall p.4)} \]

This test statistic has approximately a \[ N(0,1) \] distribution if \( H_0 \) true.

Then:

\[ N(0,1) \approx \text{approx. H0 distribution of test statistic} \]

\[ \text{Observed value: } t \]

P-value indicated by shaded area:

\[ = P\left(|N(0,1)| > 1.13\right) \]

\[ = P\left(N(0,1) < -1.13\right) + P\left(N(0,1) > 1.13\right) \]

\[ = 1 - P(-1.13 \leq N(0,1) \leq 1.13) \]

\[ = 1 - 0.74 \quad \text{(Table, p.A-104)} \]

\[ = 0.26 \]

Q. Does this prove that \( H_0 \) is true?

A. No. We have only concluded that this data is not strong evidence against \( H_0 \).

Interpretation: If \( H_0 \) true (i.e., if the coin is fair), then this result is not unusual. The chance is 26% that a more unusual (extreme) result will occur if the experiment is repeated. So we accept \( H_0 \).
(16) Average years of schooling in Town A

\( H_0: \mu = 12.3 \) versus \( H_A: \mu < 12.3 \) (one-sided)

\[
\text{test statistic} = \frac{\bar{x} - \mu_0}{\frac{SD}{\sqrt{n}}} = \cdots = -6.9
\]

(recall p.5)

\[\bar{x} = 11.6, \ SD = 4.1, \ n = 400\]

\[\text{Using data, we find that the test statistic is -6.9.}\]

This test statistic has

approximately a \( N(0,1) \)

distribution if \( H_0 \) true

Then:

\[N(0,1)\]

\[\text{Observed value of test statistic} \]

\[\text{Consistent with direction indicated in } H_A\]

\[\text{P-VALUE} = 0.0003\]

\[\text{Observed value of test statistic} \]

P-VALUE APPROXIMATELY 0.
The value -6.9 is extremely unusual. Negligible chance of a more extreme value in this direction. So this value and all more extreme values are very unlikely if \( H_0 \) is true. Conclusion: this is evidence against \( H_0 \). The hypothesis \( H_0 \) does not well explain the observed data. So we reject \( H_0 \).

Q. Does this prove that \( H_0 \) is false?
A. No. We have only established that if \( H_0 \) is true, then an event of extremely low probability has actually occurred. In short, assuming \( H_0 \) true is a bad "gamble."
Setting a Specific Threshold for "P-Value Small"

In the 2 previous illustrations, the P-value was either very large (P = 0.26) or very small (P = 0). It was straightforward to make a decision about H₀.

But we can ask: what is the borderline or threshold \( \alpha \) for P-value that corresponds to accept H₀ versus reject H₀?

A. This is a matter of choice and depends on the context. Typical choices are

- \( \alpha = 0.05 \)
- \( \alpha = 0.01 \)

When we set such a threshold \( \alpha \), then we call it the "significance level" and say that we are "conducting a test of H₀ at significance level \( \alpha \)." Then the test corresponds to the rule:

- **Accept H₀ if P-value > \( \alpha \)**
- **Reject H₀ if P-value ≤ \( \alpha \)**

It is standard to report both the decision and the P-value. The P-value indicates how strongly the evidence corresponds to the decision.

One-sided example:

\[ \begin{array}{c}
1 - \alpha \\
\hline \\
\alpha
\end{array} \]

\[ \text{Accept H₀ if test statistic here} \iff \text{Reject H₀ if test statistic here} \]
Practical Guide for Hypothesis Test about Population Mean Using $\bar{X}$

Setting

- A parent distribution with mean $\mu$ and variance $\sigma^2$, with both parameters unknown.
- Target parameter: $\mu$.
- A random sample $X_1, \ldots, X_n$ from the parent distribution.

The Point Estimator

The point estimator is

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$ 

*Key property.* By the Central Limit Theorem, the distribution of $\bar{X}$ is approximately $N\left(\mu, \frac{\sigma^2}{n}\right)$. This holds exactly if the parent distribution is Normal.

Test of $H_0 : \mu = \mu_0$

(a) Default test statistic:

$$T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}.$$ 

with $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$, the sample variance.

(b) Approximate $H_0$-distribution of $T$:

(i) default choice: $N(0, 1)$

(ii) Modification for small $n$. If $n \leq 30$, then we should allow for more uncertainty due to the variability in $s$ as estimator of $\sigma$ and due to the lessened accuracy of the CLT. For this purpose, we may replace $N(0, 1)$ by the $t_{n-1}$ distribution.

(c) Get the $p$-value using (b) (i) or (b) (ii).

(d) Interpret lower $p$-value as stronger evidence against $H_0$.

(e) For a test of $H_0$ at significance level $\alpha$, reject $H_0$ if $p$-value $\leq \alpha$ and otherwise accept $H_0$.

- RJS, 3/11/2011
Possible Errors and Their Consequences

- In estimation, the "error" is the difference between the estimator and the target parameter.
- In hypothesis testing, the "error" is a wrong decision (if any).

There are two types of such errors:

- **Type I Error**: Reject Ho when it is true
- **Type II Error**: Accept Ho when it is false

- The associated consequences can be very disparate.

**Example: Jurisprudence**

- **Type I Error**: Find an innocent person "guilty"
  
  Consequence: Incarceration of an innocent person.

- **Type II Error**: Find a guilty person "innocent"

  Consequence: A guilty person goes free to commit more crimes.

**Example: Average Years of Schooling in Town A**

- **Type I Error**: Decide mean years is below national average when really not

  Consequence: Spend unnecessary $ to strengthen educational opportunities.

- **Type II Error**: Decide mean years is not below national average when it really is.

  Consequence: Do not improve quality of educational opportunities although needed.
Can we guarantee that no error will occur?

No. We cannot eliminate the possibility of error.

Why? Because the data is inconclusive. It only provides a degree of evidence for or against Ho. We interpret the data using a test statistic which summarizes the information in the data.

The corresponding P-value measures how surprised we are by the observed test statistic value, should Ho be true.

However, we can exercise some control over the probabilities of Type I and Type II errors.

1. When we choose a significance level \( \alpha \), we are setting our tolerance for Type I error:

\[
\text{significance level } \alpha = \text{Type I Error Probability}
\]

2. A decrease in Type I error probability by modifying the test procedure is accompanied by an increase in Type II error probability, and vice versa.

Type I and Type II error probabilities trade off against each other.

3. The way to get both Type I and Type II error probabilities to be acceptably small is to choose the sample size large enough.