FALL 2013
Stat 8003: Statistical Methods I
Lecture 8

Jichun Xie

1 Hypothesis Testing

1.1 Introduction and Definition

1. Types of hypotheses
2. Types of errors
3. Choice of statistics
   - Neyman-Pearson Lemma (LRT's)
   - Likelihood-based TeSts: Generalize LRT's, Wald Tests, Score Tests
   - Tests under normality
   - Rank-based, non-parametric tests

Types of hypotheses

- Null hypothesis (H_0) status quo
- Alternative hypothesis (H_1 or H_A) – what we want to demonstrate.

Example: Clinical Trial, $\tau = \mathbb{P}$(success for surgical procedrue)

$$\mathrm{H}_0 : \tau \leq .2$$
$$\mathrm{H}_1 : \tau > .2$$

Both H_0 and H_1 are *composite* in the sense of each comprising more thatn one distribution.

Example: Incidence of West Nile Vrius. In 2002 there were serveral conformed cases of West Nile Virus in the New York metropolitan area. Let $\tau =$ the number of West Nile Virus in New York Metropolitan area per million residents

$$\mathrm{H}_0 : \tau = 55$$
$$\mathrm{H}_1 : \tau \neq 55.$$

Here H_0 is *simple*; H_1 is *composite*.

Types of Errors An hypothesis test is a data driven rule invovling a test statistic $T(\mathbf{Y})$ wehre $\mathbf{Y}$ is our data vector. We reject H_0 when $T(\mathbf{Y}) \in C$ where $C = \{\mathbf{t}\}$ is a set of possible values for $T(\mathbf{Y})$. There are two possible errors:

Decision	Truth	
	H_0 true	H_1 true
Accept H_0	$1-\alpha$	Type II error β
Reject H_0	Type I error α	$1-\beta$

- Type I error: α, or significance level of size
  - Simple: $\alpha = \mathtt{P}(\text{Reject } \mathrm{H}_0 \mid \mathrm{H}_0 \text{ true})$.
  - Composite: $\alpha = \max_{P_0 \in \mathrm{H}_0}\{\mathtt{P}(\text{Reject } P_0 \mid P_0 \text{ true})\}$
- Type II error: β,
  - Simple: $\beta = \mathtt{P}(\text{Do not reject } \mathrm{H}_0 \mid \mathrm{H}_0 \text{ not true})$.
  - Composite (not commonly used): $\beta = \max_{P_1 \in \mathrm{H}_1}\{\mathtt{P}(\text{Do not reject } \mathrm{H}_0 \mid P_1 \text{ true})\}$
- Power: $1-\beta$ (simple hypothesis only), which can be treated as a function of the alternative values of the parameter.

Example: Clinical trial, $n = 25$, $Y = \#$ patients with successful procedure. Consider a set of simple hypotheses. $T(Y) = Y$ where Y is Binomial$(n = 25, \tau)$.

$$\mathrm{H}_0 : \tau = 0.2$$
$$\mathrm{H}_1 : \tau = 0.5$$

Arbitarily suppose we chose $C = \{Y : Y > 8\}$.

$$\alpha = \mathbb{P}\left(Y > 8 \mid \tau = 0.2\right) = \sum_{y=9}^{25} \binom{n}{y} \tau^y (1-\tau)^{n-y} = 0.047.$$

In R,

```
pbinom(8,25,0.2) = 0.973
```

Now suppose

$$\mathrm{H}_0 : \tau \leq 0.2$$
$$\mathrm{H}_1 : \tau > 0.2$$

What is

$$\alpha = \max_{\tau \leq 0.2} \mathbb{P}\left(Y > 8 \mid \tau \leq 0.2\right).$$

Now consider β. First consider $\mathrm{H}_1 : \tau = 0.5$.

$$\beta = \mathtt{P}(Y \leq 8 \mid \tau = 0.5).$$

If

$$\mathrm{H}_1 : \ \tau > 0.2,$$

then β is a function, so as $1 - \beta$, which we often draw as a *power* function.

τ	.01	.05	.1	.2	.3	.4	. 5
$\mathbb{P}\left(Y > 8\right)$	0	0	0	.05	.32	.72	.95

Table 1: The value of power $= 1 - \beta$

If H_1 is true, and $\tau = 0.3$, how likely are we to get the rigth answer in our experiment?

Remark: In this experiment, it is easy to find a test statistic. What if the test statistic is not intuitive obvious? Or if we have several candidates, how do we choose the best test statistic and the rejection region? A result that can yield insight is the Neyman-Pearson Lemma.

1.2 Likelihood Ratio Test

Lemma 1 (Neyman-Pearson Lemma)**.** *We observe* $Y_1, \ldots, Y_n$ *i.i.d.* $f_Y(y)$.

$$\mathrm{H}_0 : \tau = \tau_0$$
$$\mathrm{H}_1 : \tau = \tau_1$$

The form of the most powerful test of H_0 *versus* H_1 *is given by the rule: "Reject* H_0 *for Large Values of the Likelihood Ratio"*

$$LR = \frac{Lik(\tau_1)}{Lik(\tau_0)}$$

Consider the following table and the likelihoods that would occur under the null and alternative hypotheses. If we decided that we reject when $LR \geq 2$. When should we reject in the following table?

T(Y)	Likelihood		LR
	H$_0$	H$_1$	
1	.2	.1	
2	.3	.4	
3	.3	.1	
4	.2	.4	

Steps in Constructing an Hypothesis Test:

1. Set up the hypotheses.
2. Determine the desired Type I error rate.
3. Determine a test statistic by NP lemma or one of the methods we'll discuss.
4. Find the distribution of the test statistic under the null hypothesis.
5. find the rejection region such that the Type I error rate is satisfied.
6. Possible determine power under the alternative.
7. If data are available, evaluate the test and make a decision. Determine a p-value.

Example: $Y_1, \ldots, Y_n$ *i.i.d.* Poisson(λ), $\lambda > 0$, $Y_i > 0$.

$$\begin{aligned} \mathrm{H}_0 &: \lambda = \lambda_0 \\ \mathrm{H}_1 &: \lambda = \lambda_1, \;\; \lambda_1 > \lambda_0 \end{aligned}$$

$$\begin{aligned} LR &= \frac{\lambda_1^{\sum_{i=1}^n Y_i} \exp(-n\lambda_1)}{\lambda_0^{\sum_{i=1}^n Y_i} \exp(-n\lambda_0)} \\ &= \left(\frac{\lambda_1}{\lambda_0}\right)^{\sum_{i=1}^n Y_i} \exp(-n\lambda_1 + n\lambda_0) \end{aligned}$$

Note that λ_1 and λ_0 are fixed by the definition of the hypotheses. All quantities are positive. What does the Neyman Pearson Lemma say about a rule ofr rejecting the null hypothesis? Suppose we choose to ject when $\sum_{i=1}^n Y_i$ is large? howe can we choose a cutoff such that

$$\mathbb{P}\left(\sum_{i=1}^n Y_i > c \mid \mathrm{H}_0 \text{ true}\right) = 0.05 = \alpha.$$

Approach 1: $\sum_{i=1}^{n} Y_i$ has a Poisson($n\lambda_0$) distribution if H_0 is true. (Exact result)

Approach 2: $\bar{Y}$ has an approximately Normal($\lambda_0, \lambda_0/n$) distribution. Why?

Now let's see a numerical examples.

Suppose $n = 10$, $\lambda_0 = 5$.

Approach 1: For $\alpha = 0.05$, we want to choose c such that

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i > c \mid \lambda_0 = 5\right) = 0.05.$$

or equivalently

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \leq c \mid \lambda_0 = 5\right) = 0.95.$$

$\sum_{i=1}^{n} Y_i$ has a Poisson(50) distribution under the null. In R use

```
qpois(.95,50)
```

We get the answer is 62. Then, choose $c = 62$; or if $\bar{Y}$ is used, $c = 6.2$.

Approach 2: Normal approximation. Under H_0,

$$\begin{aligned}\mathbb{P}(\bar{Y} > c) &= 0.05\\ \mathbb{P}(\bar{Y} > \lambda_0 + 1.64\sqrt{\lambda_0/n}) &= 0.05\\ c &= \lambda_0 + 1.64\sqrt{\lambda_0/n}\\ c &= 6.1596\end{aligned}$$

1.3 Generalized Likelihood Ratio Tests (GLRT)

Neyman & Pearson gave us a framework for test statistic construction when our null and alternative are simple and we have parametric distributions. In practice, hypotheses are generally composite.

1. Observe $\mathbf{Y}$ form $f_Y(y, \theta)$. Let Θ_0 and Θ_1 denote subsets of the parameter space. The GLRT rejects H_0 when

$$LR = \frac{\max_{\theta \in \Theta_1} L(\theta)}{\max_{\theta \in \Theta_0} L(\theta)} > k.$$

or a form that is a little easier to work with:

$$\Lambda = \frac{\max_{\theta \in \Theta_0 \cup \Theta_1} L(\theta)}{\max_{\theta \in \Theta_0} L(\theta)} > k^*.$$

The next problem is that we need to find the distribution of Λ in order to set up probability statements and get the error rates. Sometimes we can find an exact distribution of Λ in order to set up probability statements and get the error rates. In other cases we work with an asymptotic approximation.

2. Results which you will prove in STAT 8001 or 8002.

   1. Simple null, *e.g.*:

$$\begin{aligned} &\mathrm{H}_0 : \tau = \tau_0 \\ &\mathrm{H}_1 : \tau \neq \tau_0 \;\text{ or }\; \mathrm{H}_1 : \;\tau = \tau_1, \end{aligned}$$

   where τ is a one-dimensional parameter. Then,

$$2 \log \Lambda \,\dot{\sim}\, \chi^2(1) \text{ under } \mathrm{H}_0.$$

   2. Nested null and alternative

$$\begin{aligned} &\mathrm{H}_0 : \boldsymbol{\theta} \in \Theta_{p-r} \text{ (reduced model)} \\ &\mathrm{H}_1 : \boldsymbol{\theta} \in \Theta_p \text{ (full model)}, \end{aligned}$$

   where Θ_{p-r} is a subset of Θ_p. Then

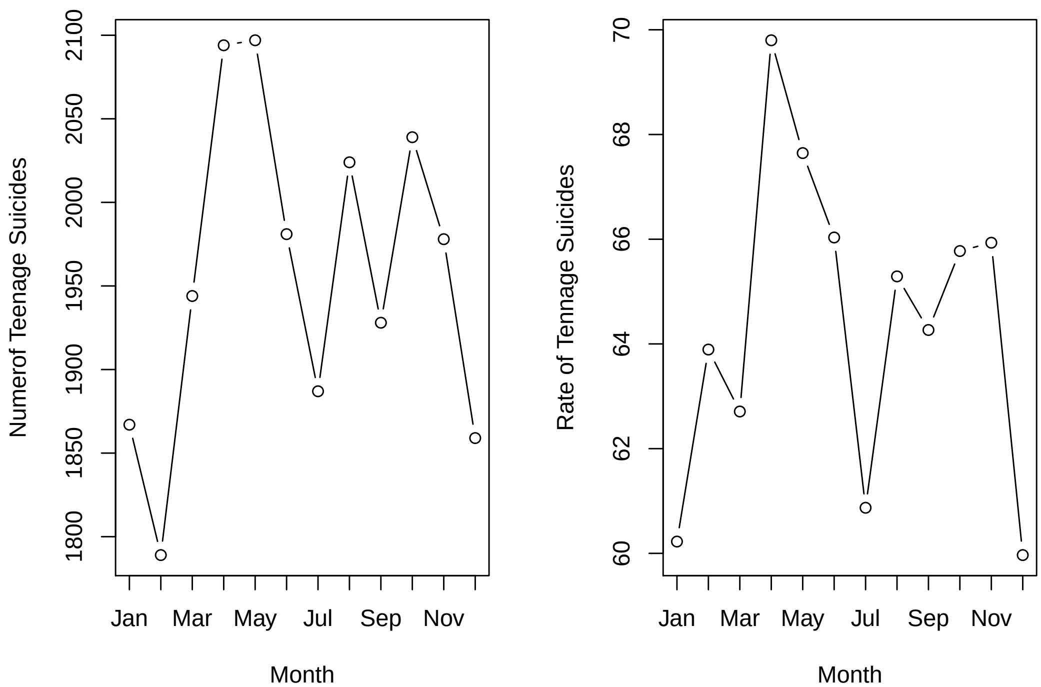
$$2 \log \Lambda \,\dot{\sim}\, \chi^2(r).$$

Examples: Seasonal Changes in Teen Suicides. For this example we need a little background information on the multinomial distribution. Let $\mathbf{Y}' = (Y_1, \ldots, Y_n)$ be a vector denoting the number of times that an independent observation falls into the ith category, $i = 1, \ldots, n$ in a series of n trials, where the probability of falling intro the ith category is θ_i; $\sum_{i=1}^n \theta_i = 1$. Then,

$$\mathbb{P}\left(Y_1 = y_1, \ldots, Y_n = y_n\right) = \frac{n!}{y_1! \cdots y_n!} \prod_{i=1}^n \theta_i^{y_i}.$$

Note that in this model, the constraint on the parameters is $\sum_{i=1}^n \theta_i = 1$; and the constraint on the counts is $\sum_{i=1}^n y_i = n$. Note that this distribution is an extension of the binomial, and the mle for each $\hat{\theta}_i = \frac{Y_i}{n}$.

Teenage Suicide Data (`http://www.familyfirstaid.org/suicide.html`). In 2001, teen suicide was the 3rd leading cause of death among young adults and adolescents 15 to 24 years of age, following unintentional injuries and homicide. The rate was 9.9/100,000 or .01%. The adolescent suicide rate among youth ages 10-14 was 1.3/100,000 or 272 deaths among 20,910,440 children in this age group. The gender ratio for this age group was 3:1 (males: females). The teen suicide rate among youth aged 15-19 was 7.9/100,000 or 1,611 deaths among 20,271,312 teenagers in this age group. The gender ratio for teenage group



was 5:1 (males: females). Among young people 20 to 24 years of age, the youth suicide rate was 12/100,000 or 2,360 deaths among 19,711,423 people in this age group. The gender ratio for this age group was 7:1 (males: females).

1. The null hypothesis is that the suicide rate is constant across months. The alternative is that there is seasonal variation in suicide rate.

2. Model: Y_i is the number of suicides in the ith month ($i = 1, \ldots, 12$).

$$\mathrm{H}_0 : \boldsymbol{\theta} = (\theta_{1,0}, \ldots, \theta_{12,0}) \text{ (reduced model)}$$
$$\mathrm{H}_1 : \boldsymbol{\theta} = (\theta_1^*, \ldots, \theta_{12}^*) \text{ (full model)}$$

where under the null hypothesis the $\theta_{i,0}$ are determined by # days in the ith month/365.

The model is

$$
\begin{aligned}
\Lambda &= \frac{\max L(\hat{\boldsymbol{\theta}}_{mle})}{\max L(\hat{\boldsymbol{\theta}}_{0,mle})} \\
&= \frac{\frac{n!}{y_1 \cdots y_n} \prod_{i=1}^{12} \hat{\theta}_i^{y_i}}{\frac{n!}{y_1 \cdots y_n} \prod_{i=1}^{12} \theta_{i,0}^{y_i}} \\
&= \prod_{i=1}^{12} \left(\frac{\hat{\theta}_i}{\theta_{i,0}} \right)^{y_i} \\
&= \prod_{i=1}^{12} \left(\frac{y_i}{n\theta_{i,0}} \right)^{y_i}.
\end{aligned}
$$

In this case, large values of Λ do not translate into a test statistic whose distribution is known. However,

$$2 \log \Lambda \dot{\sim} \chi^2(11).$$

There are $12 - 1 = 11$ free parameters in the full model since $\sum_{i=1}^{n} \theta_{i,0} = 1$; and there are 0 free parameters in the restricted model (rate determined by the number of days in the month).

$$2 \log \Lambda = 2 \sum_{i=1}^{12} y_i \left[\log(y_i) - \log(n\theta_{i,0}) \right]$$

This statistic has an "observed-expected" look to it. Why?

See R handout for the results of the example. $2 \log \Lambda$ is 47.66 based on the data; and the $\alpha = 0.05$ level cut-off is 19.67. So we should reject the H_0 under the level of 0.05. That is to say, there is a significant difference among the teenage suicide rate per month.

Lastly Collaborators often like to report p-values in their work. We can think of a p-value as the probability of the observed result, or a more extreme result, given the null hypothesis is true. Note: A p-value is a random variable (it's a function of the data) whereas the level of the test is fixed by the design. For the suicide data the p-value is $p = 1.7 \times 10^{-6}$ or $p < .0001$.

1.4 Other likelihood-based hypothesis tests

For the hypotheses:

$$
\begin{aligned}
&\mathrm{H}_0 : \theta = \theta_0 \\
&\mathrm{H}_1 : \theta \neq \theta_0.
\end{aligned}
$$

1.4.1 Wald Test

Recall that we discussed in the previous lecture the variance of the MLE estimators. Suppose $Y_1, \ldots, Y_n \sim f(y;\theta)$, *i.i.d.*. We define the log likelihood function by

$$l(\theta) = \sum_{i=1}^{n} \log f(y_i;\theta).$$

The MLE $\hat{\theta}_{mle}$ maximized the (log) likelihood:

$$\hat{\theta}_{mle} = \arg\max_{\theta} l(\theta).$$

Suppose the Fisher information

$$I(\theta) = \mathtt{E}\left\{ (l'(\theta))^2 \middle| \, \theta \right\} = -\mathtt{E}\left\{ l''(\theta) | \, \theta \right\}$$

exists. Then under some regularity conditions,

$$\mathtt{Var}(\hat{\theta}_{mle}) = (I(\theta))^{-1}.$$

Now define the observed Fisher information

$$i(\hat{\theta}) = I(\hat{\theta}).$$

Then we can estimate the variance of MLE by

$$\widehat{\mathtt{Var}}(\hat{\theta}_{mle}) = i^{-1}(\hat{\theta}_{mle}).$$

To test $\mathrm{H}_0: \ \theta = \theta_0$, we construct

$$W = \frac{|\hat{\theta}_{mle} - \theta_0|}{\sqrt{i(\hat{\theta}_{mle})^{-1}}},$$

Under the null hypothesis, W has a standard normal distribution. Can do one-sided tests as well.

Example. For Poisson distribution,

$$W = \frac{\sqrt{n}|\bar{Y} - \tau_0|}{\sqrt{\bar{Y}}}.$$

More generally if we have an estimate of $\hat{\theta}$ of θ that is asymptotically normal, you sometimes see a 'Wald-type' test

$$W = \frac{|\hat{\theta} - \theta_0|}{\sqrt{\mathtt{Var}(\hat{\theta})}},$$

where $\mathtt{Var}(\hat{\theta})$ is the variance of $\hat{\theta}$. We generally use a consistent estimate of $\mathtt{Var}(\hat{\theta})$ and justify the asymptotic distribution of W using Slutsky's theorem.

1.4.2 Score Test

The score is

$$U(\theta) = \frac{\mathrm{d}l(\theta)}{\mathrm{d}\theta},$$

where under H_0,

$$\mathtt{E}[U(\theta_0)] = 0 \text{ and } \mathtt{Var}[U(\theta_0)] = I(\theta_0),$$

where $I(\theta_0)$ is the information evaluated at θ_0.

Now $U(\theta)$ is the sum of *i.i.d.* random variables (since the log of the likelihood is the sum of the log likelihood for the individual observations). Thus by the central limit theorem

$$\frac{U(\theta_0)}{\sqrt{I(\theta_0)}} \xrightarrow{D} N(0,1).$$

The beauty of the score is that it does not require finding the mle, and thus it sometimes taks a simple form.

Example. For the Poisson recall that

$$\begin{aligned} f_Y(y) &= \theta^y \exp(-\theta)/y! \\ U(\theta) &= -n + \frac{\sum_{i=1}^n Y_i}{\theta} \end{aligned}$$

$$\begin{aligned} \frac{\mathrm{d}^2 \log L(\theta, \mathbf{y})}{\mathrm{d}\theta^2} &= \frac{-\sum_{i=1}^n Y_i}{\theta^2} \\ I(\theta_0) &= \left.\frac{n}{\theta}\right|_{\theta=\theta_0} \end{aligned}$$

$$\begin{aligned} U &= \frac{-n + \frac{\sum_{i=1}^n Y_i}{\theta_0}}{\sqrt{\frac{n}{\theta_0}}} \\ &= \sqrt{\frac{n}{\theta_0}}(\bar{Y} - \theta_0) \end{aligned}$$

Questions:

- How to compare different methods?
- What will happen if the null hypothesis is true and the sample size is large?
- What will happen if the alternative hypothesis is true and the sample size is large?