

# Statistics 312: Introduction To Mathematical Statistics Lecture 15 (Review)

PROFESSOR MICHAEL R. KOSOROK  
DEPARTMENT OF STATISTICS

## Outline

1. Graphical and Numerical Summaries
2. The Normal Curve
3. Populations and samples
4. Confidence Intervals
5. Dichotomous Data
6. The Bootstrap
7. Stratified Random Sampling
8. The Lognormal Distribution
9. Method of Moments Estimation
10. Maximum Likelihood Estimation
11. Hypothesis Testing

1

12. Confidence Intervals and Hypothesis Tests
13. Generalized Likelihood Ratio Tests
14. Goodness-of-Fit Tests
15. Two-Way Contingency Tables
16. Two-Sample t-Tests
17. The Mann-Whitney-Wilcoxon Test
18. The paired t-Test
19. The Signed Rank Test
20. The Test

2

## 1 Graphical and Numerical Summaries

- Histogram
- Box-and-Whisker Plots
- Quantile-Quantile Plots
- Bar Charts
- Empirical Cumulative Distribution Functions
- Mean
- Standard deviation
- Median
- Other percentiles
- Interquartile range (IQR)
- Five-Number Summary
- Skewness
- Kurtosis
- Cross-Tabulations (Contingency Tables)

3

## 2 The Normal Curve

The normal curve is a density for continuous random quantities which mimics certain natural phenomena. A density is an idealized histogram, where the probability for an interval is equal to the area under the curve over the interval.

The curve is unimodal, symmetric around 0, and follows the 68–95–99.7 rule. The formula for the density is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

The area under the curve to the left of  $x$  is  $\Phi(x) = \int_0^x \phi(u) du$ .

The mean and median for the normal curve is 0, the standard deviation is 1, the skewness is 0, and the kurtosis is 3. Comparing skewness and kurtosis for an arbitrary data set to 0 and 3, respectively, is one way of assessing “normality.”

4

### 3 Populations and samples

- Population units
- Population size
- Unit characteristic
- Population parameter
- Sample units
- Sample size
- Sample statistics
- Selection rule
  - Without replacement
  - With replacement
  - Representative sampling

5

### 5 Dichotomous Data

Dichotomous data have only the values 0 or 1 and are used to indicate the presence or absence of something (such as whether someone owns a PC).

In this setting,  $\pi = \sum x_i/N$  is the mean but also the proportion of 1's. The variance simplifies to  $\sigma^2 = \pi(1 - \pi)$ . It is also easy to derive that  $s^2 = \bar{x}(1 - \bar{x})n/(n - 1)$ , and thus an unbiased estimate of  $\text{var}[\bar{x}]$  is

$$\frac{\bar{x}(1 - \bar{x})}{n - 1} \times \frac{N - n}{N}.$$

7

### 4 Confidence Intervals

For a sample of size  $n$  (simple random),  $x_1, \dots, x_n$ , a good estimator for  $\sigma^2$  is

$$s^2 = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

An unbiased estimated of  $\text{var}[\bar{x}]$  is

$$\frac{s^2}{n} \times \frac{N - n}{N},$$

which becomes  $s^2/n$  for large  $N$ .

A  $100(1 - \alpha)\%$  confidence interval for  $\mu$  can be based on the fact that, for normal data,

$$T = \frac{\sqrt{n}(\bar{x} - \mu)}{s}$$

is t-distributed with  $n - 1$  degrees of freedom. Thus a good two-sided confidence interval is

$$\bar{x} \pm t_{1-\alpha/2}(n - 1) \frac{s}{\sqrt{n}}.$$

6

### 6 The Bootstrap

A bootstrap,  $T^*$ , of a statistic  $T$  is obtained by selecting a random sample of size  $n$  with replacement from the original sample.

This procedure is repeated many times so that the (re)sampling distribution of  $\bar{x}^*$  can be obtained. Confidence intervals can be based on this resampled quantity by taking the 2.5%-ile and the 97.5%-ile of the resampled  $\bar{x}^*$ -s. This procedure is called bootstrapping and works extremely well.

The parametric bootstrap is a variant of the bootstrap which takes advantage of parametric assumptions about the distribution involved.

8

## 7 Stratified Random Sampling

A stratified random sample is a probability method of sampling that involves dividing a population into appropriate strata (subgroups) and then taking a simple random sample from each strata. The Minnesota radon data was an example of this concept.

How can the sample sizes  $n_j$  be chosen to minimize the variance of  $\tilde{x}$ ? There are two main approaches:

- **Optimal Allocation:** this requires some knowledge of the within-strata variances, but, if such knowledge is available, it is the best approach. It decrees that  $n_j/n$  be as close to

$$\frac{w_j \sigma_j}{\sum_{j=1}^J w_j \sigma_j}$$

as possible, where  $w_j = N_j/N$ .

- **Proportional Allocation:** decrees that  $n_j/n$  be as close to  $w_j$  as possible.

9

### 8.1 Method of Moments Estimation

For a family of densities  $f(x|\theta)$ ,  $\theta \in \Theta$ , the  $k$ 'th moment is

- $m_k = \sum_{x \in \mathcal{X}} x^k f(x|\theta)$  for discrete RV's, or
- $m_k = \int_{\mathcal{X}} x^k f(x|\theta) dx$  for continuous RV's.

The idea is to compute as many moments as the dimension of  $\theta$ , match the moments with sample moments, and solve for  $\theta$ .

11

## 8 The Lognormal Distribution

Proportions in the world seem to follow the lognormal distribution.

If  $Y = \log X$  has a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , then  $X$  is lognormal distributed.

The density for  $X$ , via the change of variable formula for the transformation  $X = \exp\{Y\}$  is

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} \left(\log \frac{x}{\gamma}\right)^2\right],$$

where  $\gamma = \exp(\mu)$  is the geometric mean of  $X$ .

10

### 8.2 Maximum Likelihood Estimation

The main quantities to remember:

1.  $\ell_n(\theta) = \sum_{i=1}^n \log f(x_i|\theta)$  (the log-likelihood)
2.  $S_n(\theta) = \frac{\partial \ell_n(\theta)}{\partial \theta}$  (the "score function")
3.  $I_n(\theta) = -\frac{\partial S_n(\theta)}{\partial \theta}$  (the "observed information")
4.  $I(\theta) = E_{\theta} [n^{-1} I_n(\theta)]$  (the "information")
5.  $\hat{\theta}_n = \arg \max_{\theta \in \Theta} \ell_n(\theta)$  (the MLE)

The main points:

1.  $\sqrt{nI(\theta_0)}(\hat{\theta}_n - \theta)$  is approximately standard normal.
2.  $I(\theta_0)$  can be consistently estimated by  $\hat{\sigma}_n^2 = n/I_n(\hat{\theta}_n)$ .
3. An approximate 95% confidence interval is  $\hat{\theta}_n \pm 2\hat{\sigma}_n/\sqrt{n}$ .

12

## 9 Hypothesis Testing

Statistical hypothesis testing is a formal way of distinguishing between competing probability models:

- Null hypothesis
- Alternative hypothesis
- Simple hypotheses
- Composite hypotheses

Consider the problem of testing  $H_0 : \mu = 0$  versus some alternative. Two common composite alternatives are

- *One-sided* such as  $H_A : \mu > 0$  or
- *Two-sided* such as  $H_A : \mu \neq 0$ .

Is the chisquare test of independence one-sided or two-sided?

13

The probability distribution of the test statistics under  $H_0$  is called the *null distribution*.

The *p-value* for a specific value of the statistic  $T = t$ , is the smallest type I error resulting in rejecting  $H_0$  if  $T \geq t$  (one-sided p-value =  $P(T \geq t|H_0)$  or  $|T| > t$  (two-sided p-value =  $P(|T| \geq t|H_0)$ ). *More precisely, it is the probability, under  $H_0$ , of the statistic  $T$  being at least as extreme as  $t$ , where "extreme" means having evidence against the null distribution.*

15

A *type I error* is made when  $H_0$  is true but  $T$  is in the rejection region. The probability of making a type I error under  $H_0$  is called the *significance level* or *size* of the test, and is often denoted  $\alpha$ .

A *type II error* is made when  $H_A$  is true but  $T$  is in the acceptance region. The probability that  $H_0$  is rejected when  $H_A$  is true ( $H_0$  is false) is the *power*, and is  $1 - \beta$ , where  $\beta$  is the probability of a type II error.

14

## 10 Confidence Intervals and Hypothesis Tests

Let  $X_1, \dots, X_n$  by i.i.d.  $N(\mu, \sigma^2)$ , where  $\mu$  is unknown. Let  $T = \bar{x}$ . Then  $\sqrt{n}(\bar{T} - \mu) \sim N(0, \sigma^2)$ .

A  $100(1 - \alpha)\%$  confidence interval for  $\mu$  is  $T \pm t_{1-\alpha/2}(n-1) \times s/\sqrt{n}$ , where  $t_p(\nu)$  is the  $p$ -th quantile of a t distribution with  $\nu$  degrees of freedom and  $s^2$  is the sample variance (with  $n - 1$  in the denominator).

16

Suppose we wish to test  $H_0 : \mu = \mu_0$  versus  $H_A : \mu \neq \mu_0$  for some specific  $\mu_0$  (often  $\mu_0 = 0$ ).

If we use a two-sided critical region, where we reject  $H_0$  if  $|\bar{x} - \mu_0| > t$ , then choosing  $t = t_{1-\alpha/2}(n-1)s/\sqrt{n}$  will result in a test of size  $\alpha$ .

Hence if  $\bar{x} \pm t$  contains  $\mu_0$ , we accept  $H_0 : \mu = \mu_0$ . Thus, a confidence interval contains those values of  $\mu_0$  for which the test statistic  $T = \bar{x}$  would accept  $H_0 : \mu = \mu_0$ .

## 11 Generalized Likelihood Ratio Tests

Let the sample  $X_1, \dots, X_n$  be i.i.d. with density family  $f(x|\theta)$ ,  $\theta \in \Theta$ . Let  $\Theta_0$  be a subset of  $\Theta$  with fewer free parameters, and denote

$$L_n(\theta) = \prod_{i=1}^n f(x_i|\theta).$$

The *generalized likelihood ratio* is:

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} L_n(\theta)}{\sup_{\theta \in \Theta} L_n(\theta)}.$$

(sup means to take the maximum.)

The *generalized likelihood ratio test* rejects  $H_0 : \theta \in \Theta_0$  versus  $H_A : \theta \notin \Theta_0$  for small values of  $\Lambda$ , say  $\Lambda \leq \lambda_0$ , where  $\lambda_0$  is chosen to ensure a type I error of  $\alpha$ .

What about large values of  $\Lambda$ ?

17

Usually  $\Theta_0$  consists of special cases of the model  $\Theta$ . For example, consider the normal setting, where

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}.$$

While  $\Theta$  consists of all values of  $\mu$  and  $\sigma^2 \geq 0$ , Let  $\Theta_0$  consist of all values of  $(\mu, \sigma^2)$  satisfying  $\mu = 0$  and leaving  $\sigma^2$  unspecified.

In this setting,  $\Theta$  is two-dimensional while  $\Theta_0$  is one-dimensional. In settings like this,  $-2 \log \Lambda$  is approximately chi-square distributed with degrees of freedom equal to  $d = \dim(\Theta) - \dim(\Theta_0)$  (which for the normal example above is 1).

Thus one could look up the  $1 - \alpha$  quantile  $\chi_{1-\alpha}$  of a chi-square distribution with  $d$  degrees of freedom, and reject  $H_0 : \theta \in \Theta_0$  in favor of  $H_A : \theta \notin \Theta_0$  if  $-2 \log \Lambda \geq \chi_{1-\alpha}$ .

19

18

## 12 Goodness-of-Fit Tests

Goodness-of-Fit tests are used to evaluate whether a sample comes from a hypothesized distribution. These tests are similar to the multinomial generalized likelihood ratio tests.

Here is the procedure:

- The sample space is divided into  $K$  bins (not too unevenly spaced).
- The probability  $p_k(\theta)$  of falling into the  $k$ 'th bin is estimated using the hypothesize distribution, to obtain  $p_k(\hat{\theta})$ , and the expected number  $E_k = np_k(\hat{\theta})$  is computed,  $k = 1 \dots K$ .

20

- The actual number  $O_k$  from the sample falling into the  $k$ 'th bin is computed,  $k = 1 \dots K$ .
- The statistic

$$T = \sum_{k=1}^K \frac{(O_k - E_k)^2}{E_k}$$

is calculated.

Under the null hypothesis that the samples comes from the hypothesized distribution, and provided certain regularity conditions hold,  $T$  is approximately chi-square distributed with degrees of freedom  $K - 1 - d$ , where  $d$  is the number of free parameters in the hypothesized model.

The Kolmogorov-Smirnov (KS) test is another method of testing whether the data has a certain distribution. An advantage of the KS test over the goodness-of-fit test is that bins and bin sizes do not have to be selected.

## 14 Two-Way Contingency Tables

The data for an  $r \times c$  two-way contingency table (with  $r$  rows and  $c$  columns) has the general form

$i$	$j$				total
	1	2	$\dots$	$c$	
1	$n_{11}$	$n_{12}$	$\dots$	$n_{1c}$	$n_{1\cdot}$
2	$n_{21}$	$n_{22}$	$\dots$	$n_{2c}$	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
$r$	$n_{r1}$	$n_{r2}$	$\dots$	$n_{rc}$	$n_{r\cdot}$
total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot c}$	$n_{\cdot\cdot}$

## 13 Randomized Experiments

What does it mean to randomize an experiment?

When can Fisher's exact test be used?

What are the dangers of not randomizing in a clinical trial?

What is the relationship between Fisher's exact test and the hypergeometric distribution?

The null hypothesis is that the column variable is statistically independent of the row variable.

In both cases, this null hypothesis is also tested with the same statistic

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - n_i n_{\cdot j} / n_{\cdot\cdot})^2}{n_i n_{\cdot j} / n_{\cdot\cdot}},$$

which is chi-squared with  $(r-1)(c-1)$  degrees of freedom under the null hypothesis.

Even though we typically use only the right tail of the chi-squared distribution to compute p-values, the test is not one-sided per se (the alternative hypothesis is global).

The  $n_{ij}$  terms are referred to as *observed values* while the  $n_i n_{\cdot j} / n_{\cdot\cdot}$  terms are referred to as *expected values* (under the null hypothesis).

## 15 Two-Sample t-Tests

Suppose we have two independent normal samples,  $X_1, \dots, X_n$  and  $Y_1, \dots, Y_m$ , with corresponding means  $\mu_x$  and  $\mu_y$  and common variance  $\sigma^2$ .

Since we don't know  $\sigma^2$ , we estimate it with the *pooled sample variance*

$$s_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{m+n-2}.$$

$s_p^2$  is the total squared deviations from the means divided by the total degrees of freedom (2 are taken off because 2 means were estimated).

The statistic

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

has a t distribution with  $m+n-2$  degrees of freedom, and a  $100(1-\alpha)\%$  confidence interval for  $\mu_x - \mu_y$  is  $(\bar{X} - \bar{Y}) \pm t_{1-\alpha/2}(m+n-2)s_p \sqrt{\frac{1}{n} + \frac{1}{m}}$ .

25

It can be shown that when there are no ties,

$$E[R] = \frac{m(m+n+1)}{2}$$

and

$$\text{var}[R] = \frac{mn(m+n+1)}{12}.$$

Furthermore, when  $m$  and  $n$  are large,

$$Z = \frac{R - E[R]}{\sqrt{\text{var}[R]}}$$

is approximately standard normal. This can be used to obtain p-values.

When there are ties, the p-values obtained using  $Z$  will be *conservative* estimates of the true p-value (meaning they will be over-estimates).

27

## 16 The Mann-Whitney-Wilcoxon Test

$n$  units are randomly chosen and assigned to group 1, while  $m$  units are randomly chosen and assigned to group 2.

The null hypothesis is  $H_0$ : there is no group difference.

The test statistic is calculated as follows:

- Group all  $m+n$  observations together and rank them in order.
- If there are ties, take the values that are tied, average the ranks that would have been assigned had they not been tied, and then use that average as the rank for each of the tied values.
- Add up all of the ranks associated with group 2 (sample size  $m$ ), and call this value  $R$ .
- If  $R$  is too large or too small, reject  $H_0$ .

26

## 17 The Paired t-Test

Frequently, treatment and control observations are paired rather than independent. For example, treatment may be given to a randomly selected eye while the remaining eye receives the standard treatment, or two different laboratory techniques may be applied to the same samples.

In these settings, the data is of the form  $(X_i, Y_i)$ ,  $i = 1 \dots n$ , where the  $X$  data involves one treatment while the  $Y$  data another. The pairs are independent of each other, but the values within pairs may be quite correlated.

If  $X_1$  and  $Y_1$  are independent, the two-sample methods we have just discussed are applicable. If they are not independent, two-sample methods are not valid and paired methods are needed.

28

## 18 The Signed Rank Test

In some settings, the normality assumption may not be appropriate, and a nonparametric approach would be more appealing. One such approach is the signed rank statistic.

The Wilcoxon signed rank statistic is computed as follows:

- Rank the absolute values of the differences  $|D_i|$ ,  $i = 1 \dots n$ , and denote these ranks  $D_1, \dots, D_n$ .
- Restore the signs of the  $D_i$  to the ranks, obtaining the *signed ranks*.
- Calculate  $W_+$ , the sum of those ranks that have positive sign.

When should we use the paired t-test rather than the signed rank test?

## 19 The Test

The test consists of

- 5 fill-in-the-blank (or circle-the-right-answer) questions (worth 2 points each, 10 total),
- 5 short answer questions (worth 4 points each, 20 total),
- 1 medium short question [define p-value in your own words] (worth 8 points).

- 4 long questions:
  - A confidence interval based on the t distribution (worth 10 points),
  - A maximum likelihood and method of moments problem (worth 28 points),
  - A contingency table problem (worth 20 points), and a
  - Mann-Whitney-Wilcoxon test problem [you will need to be able to handle ties, compute  $R$ , and know the consequences of p-values based on the continuity (no ties) assumption in the presence of ties; for the extra credit, you will need to know how to compute the mean, variance, and  $Z$ -statistic under the no ties assumption] (worth 18 points).

The total points possible are 114, 14 of which are extra credit.