STA 580 — Fall 2008 — Dr. Charnigo

Lecture 8

Non-normal populations and outlying observations

Introduction. Real data sets often come from non-normal populations, in which case many of the inferential procedures from Lectures 4 through 7 do not apply. Moreover, real data sets often contain outlying observations that may unduly influence conclusions reached via the inferential procedures from Lectures 4 through 7. Such outlying observations often represent errors in data entry, but they may also be genuine.

Detection of non-normality and outlying observations. An informal but effective approach for detecting non-normality and outlying observations is to inspect a graphical display. Definitions 2.11 and 2.12 provide guidelines for identifying outliers in box plots; Figures 8.11 and 8.12 illustrate outliers in box plots with the "0" symbol. If we find outlying observations that we suspect represent errors in data entry, we can try to recover the correct sample values. If we find outlying observations that are genuine, we have four options for data analysis. Options 2, 3, and 4 are also useful for addressing non-normality.

OPTION 1. Presenting two sets of results. We can analyze the data with and without the outlying observations. If the two sets of results yield conclusions that are in qualitative agreement (e.g., we reject the null hypothesis either way), then presenting both sets of results and noting the agreement may be satisfactory. In presenting two sets of results, typically we regard one set of results as part of a "primary" analysis and the other set of results as part of a "secondary" or "sensitivity" analysis. Designations of "primary" and "secondary"/"sensitivity" make reading (and citing) our work easier for other researchers.

Potential drawback of this option: Sometimes the two sets of results will yield qualitatively different conclusions.

OPTION 2. Discretization. We can discretize the response variable and then consider population proportions instead of population means. For instance, we can label finger-wrist tapping scores (Cf. Figures 8.11 and 8.12) above 50 as "high" and less than or equal to 50 as "low". Then, instead of asking whether the mean finger-wrist tapping score in the exposed population differs from that in the control population, we can ask whether the proportion of low scores in the exposed population differs from that in the control population. Since a finger-wrist tapping score of 8 is simply regarded as low, just the same as a finger-wrist tapping score of 38, outlying observations no longer have undue influence.

Potential drawback of this option: The choice of a cutoff, such as 50, is rather arbitrary. Even with a carefully chosen cutoff, the power to distinguish between populations may be much lower than desired since much information is lost in discretization.

Remark: We will discuss how to compare population proportions later in this lecture.

OPTION 3. Nonparametric techniques. We can analyze the data using nonparametric techniques, which typically assign "ranks" to the observations and then entail arithmetic calculations on the ranks rather than on the observations themselves. An outlying observation may end up being ranked first or last, but that is the limit of its ability to influence conclusions.

Potential drawback of this option: The power to distinguish between populations may be somewhat lower than desired since some information is lost in assigning ranks.

Remark: We will discuss some nonparametric techniques in Lecture 9.

OPTION 4. Nonlinear transformation. We can apply a nonlinear transformation to the response variable and then analyze the transformed data instead of the original data. The histogram in Figure 1 shows data to which we would be highly uncomfortable applying any procedure that assumed normality. However, after taking (natural) logarithms of the observations, we obtain the histogram in Figure 2. Now we would be quite comfortable with a normality assumption.

The most commonly employed transformations are $X \mapsto \log(X)$ (for positive data with a strong right skew), $X \mapsto \log(1+X)$ (for nonnegative data with a strong right skew), and $X \mapsto \sqrt{X}$ (for nonnegative data with a moderate right skew). [1]

Potential drawback of this option: Interpreting the results can be difficult. For instance, if we apply a logarithmic transformation and then reject $H_0 : \nu_1 = \nu_2$, where $\nu_1 := E[\log(X_1)]$ and $\nu_2 := E[\log(Y_1)]$, this is *not* logically equivalent to rejecting $H_0 : \mu_1 = \mu_2$, where $\mu_1 := E[X_1]$ and $\mu_2 := E[Y_1]$. [2]

[1] For those who have had calculus, the key to an effective transformation rests with the first two derivatives of the transforming function. The first derivative should be positive, to ensure that the ordering of observations is preserved by the transformation. The second derivative should be negative if the original data have a right skew, to "shrink" extremely large sample values. The second derivative should be positive if the original data have a left skew.

[2] For those who have taken a semester-long course in probability, this non-equivalence is due to Jensen's Inequality.

Figure 1:

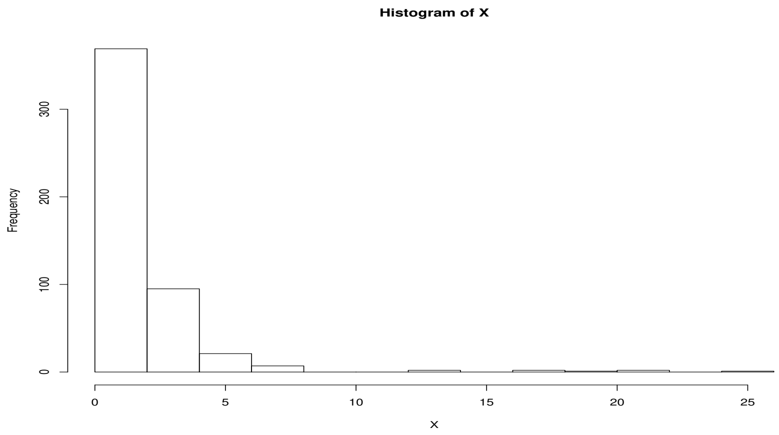
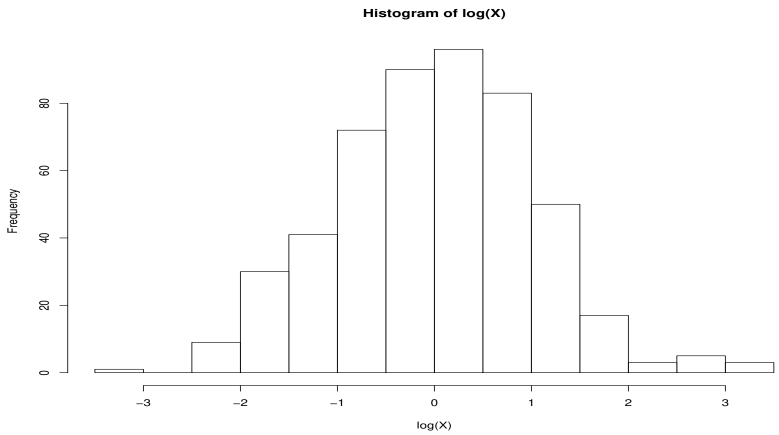


Figure 2:



Comparing proportions (independent samples)[3]

Introduction. Let p_1 denote the proportion of individuals in the first population for whom a certain statement is true, and let p_2 denote the proportion of individuals in the second population for whom the statement is true. Suppose that we want to conduct a level α test of

$$H_0 : p_1 = p_2 \quad \text{against} \quad H_1 : p_1 \neq p_2.$$

For example, p_1 may denote the fraction of nonsmoking children and adolescents with FEV scores less than 3 and p_2 the fraction of smoking children and adolescents with FEV scores less than 3.

The hypothesis test. Let $\hat{P}_1$ and $\hat{P}_2$ denote the random conceptualizations of the sample proportions. If $n_1p_1(1-p_1) \geq 5$ and $n_2p_2(1-p_2) \geq 5$, then the Central Limit Theorem implies that

$$\frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}$$

is approximately standard normal. Hence, if H_0 is true (Equation 10.1),

$$Z := \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{p(1-p)/n_1 + p(1-p)/n_2}}$$

is approximately standard normal, where p denotes the common (but unknown) value of p_1 and p_2. An estimate for p based on both samples is (Equation 10.2)

$$\hat{p} := \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

[3]Methodology for comparing proportions via paired samples is provided in Equation 10.12. That material is beyond the scope of STA 580 but is not difficult for you to read if you choose.

Thus, we obtain an approximate level α test by rejecting H_0 if

$$z := \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$$

is either greater than $z_{1-\alpha/2}$ or less than $-z_{1-\alpha/2}$ (Equation 10.3 without Rosner's continuity correction and mistaken statement about when to reject H_0). The approximate p-value is $2P(Z \leq z)$ if $z \leq 0$ and $2P(Z > z)$ if $z > 0$, where Z has a standard normal distribution and z is the numerical test statistic. An approximate $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Note that $\hat{p}$ does not appear in this expression, as we do not assume H_0 to be true when we construct the confidence interval.

In practice we cannot say whether $n_1 p_1(1-p_1) \geq 5$ and $n_2 p_2(1-p_2) \geq 5$ because p_1 and p_2 are unknown. Thus, we check whether $n_1\hat{p}_1(1-\hat{p}_1) \geq 5$ and $n_2\hat{p}_2(1-\hat{p}_2) \geq 5$.

Example (the hypothesis test). Refer to page 2 of {FEV.pdf}. Let us conduct a level $\alpha = 0.05$ test of

$$H_0 : p_1 = p_2 \quad \text{against} \quad H_1 : p_1 \neq p_2,$$

with p_1 and p_2 as defined earlier. Out of the 589 nonsmokers, 431 (73.17%) had FEV scores less than 3. Out of the 65 smokers, 20 (30.77%) had FEV scores less than 3. Out of the 451 people with FEV scores less than 3, 431 (95.57%) were nonsmokers.

Continue to page 3 of {FEV.pdf}. We have $\hat{p}_1 = 0.7317$, $\hat{p}_2 = 0.3077$, and $\hat{p}_1 - \hat{p}_2 = 0.4241$. Moreover,

$$\hat{p} = \frac{589(0.7317) + 65(0.3077)}{589 + 65} = \frac{451}{654} = 0.6896.$$

Hence,

$$z = \frac{0.4241}{\sqrt{0.6896(1-0.6896)(1/589+1/65)}} = 7.013.$$

Since $z_{0.975} = 1.96$, we reject H_0.

Approach via contingency table. A general 2×2 contingency table is organized as shown below (Table 10.7). Of course, we have already seen an example on page 2 of {FEV.pdf}.

Sample	Statement True	Statement False	Row Total
1	a	b	$a+b$
2	c	d	$c+d$
Column Total	$a+c$	$b+d$	$a+b+c+d$

Note that $\hat{p}_1 = a/(a+b)$, $\hat{p}_2 = c/(c+d)$, and $\hat{p} = (a+c)/(a+b+c+d)$.

To see how we can test $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$ via a contingency table, let us consider what would happen if H_0 were true. We would expect the statement of interest to hold for

$$n_1\hat{p} = \frac{(a+b)(a+c)}{(a+b+c+d)}$$

individuals in the first sample and for

$$n_2\hat{p} = \frac{(c+d)(a+c)}{(a+b+c+d)}$$

individuals in the second sample. Call these numbers E_{11} and E_{21}, respectively. Define $E_{12} := n_1 - E_{11}$ and $E_{22} := n_2 - E_{21}$ to be the expected numbers of individuals in the two samples for whom the statement of interest would *not* hold.

Our decision whether to reject H_0 is based on how much the *expected* values $E_{11}, E_{12}, E_{21}, E_{22}$ disagree with the *observed* values a, b, c, d. Let

$$\chi^2 := \frac{(a - E_{11})^2}{E_{11}} + \frac{(b - E_{12})^2}{E_{12}} + \frac{(c - E_{21})^2}{E_{21}} + \frac{(d - E_{22})^2}{E_{22}}.$$

This is different from Equation 10.5 only in that the Yates continuity correction (the repeated -0.5) has been omitted.

If $n_1 p_1(1 - p_1) \geq 5$ and $n_2 p_2(1 - p_2) \geq 5$, then we obtain an approximate level α test by rejecting H_0 if $\chi^2 > \chi^2_{1,1-\alpha}$. Since p_1 and p_2 are unknown, in practice we ask whether each of the expected values is at least 5. [4]

Example (approach via contingency table). Referring to page 2 of {FEV.pdf}, we have $E_{11} = \frac{(589)(451)}{654} = 406.18$, $E_{21} = \frac{(65)(451)}{654} = 44.82$, $E_{12} = 182.82$, and $E_{22} = 20.18$. So, we have

$$\chi^2 = \frac{(431 - 406.18)^2}{406.18} + \frac{(158 - 182.82)^2}{182.82} + \frac{(20 - 44.82)^2}{44.82} + \frac{(45 - 20.18)^2}{20.18} = 49.18.$$

Since $\chi^2_{1,0.95} = 3.84$, we reject H_0.[5]

Remark. The approach via contingency table generalizes to accommodate situations in which there are three or more populations being compared with respect to how their members are classified into three or more categories. This generalization is referred to as the chi-square test for association (Equations 10.22, 10.23).

[4]If the sample sizes are not large, you can employ Fisher's exact test. A detailed description is beyond the scope of STA 580, but you can review Equation 10.11 if interested. In any case, SAS automatically performs Fisher's exact test.

[5]With reasonable allowance for rounding, we have $\chi^2 = z^2$ in this example. In fact, that is a general phenomenon. Moreover, one can show that $\chi^2_{1,1-\alpha} = (z_{1-\alpha/2})^2$ for any α between 0 and 1. Hence, the z statistic and chi-square statistic approaches are equivalent.

Power and sample size for comparing means (independent samples) [6]

Introduction. We discuss power and sample size computations in the context of testing

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{against} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

based on two independent samples (Equations 8.27 and 8.28).

Notation. Let Δ be the anticipated absolute difference in means, $|\mu_1 - \mu_2|$. In the absence of better knowledge or better insight, we usually take $\Delta := |\bar{x} - \bar{y}|$, where $\bar{x}$ and $\bar{y}$ come from a pilot study. Also, we usually take s_x^2 and s_y^2 as proxies for σ_1^2 and σ_2^2. If we want to calculate power for given sample sizes n_1 and n_2, we let $k := n_2/n_1$ be the ratio of sample sizes. If we want to calculate sample sizes for given power $1 - \beta$, we let k denote the desired ratio of sample sizes.

Power. The approximate power for a level α test is (Equation 8.28)

$$\Phi\left[-z_{1-\alpha/2} + \frac{\sqrt{n_1}\,\Delta}{\sqrt{\sigma_1^2 + \sigma_2^2/k}}\right].$$

Example (power). Referring to {FEV.pdf}, let us find the power for a level $\alpha = 0.05$ test of

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{against} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

if $n_1 = 40$ and $n_2 = 20$. We have $\Delta = |2.5661 - 3.2769| = 0.7108$ and

[6] Equations 10.14 and 10.15 show how to perform power and sample size calculations for comparing proportions via independent samples. You will not be tested on that material in STA 580, but you are encouraged to note its presence in the textbook if you anticipate having use for it in the future.

$k = 20/40 = 0.50$, so that the power is

$$\Phi\left[-1.960 + \frac{\sqrt{40}\,(0.7108)}{\sqrt{0.7234 + 0.5625/0.50}}\right] = \Phi[1.347] = 0.911.$$

Sample size. To a close approximation, the sample size n_1 required to attain power $1 - \beta$ is (Equation 8.27) the smallest integer greater than or equal to

$$\frac{(\sigma_1^2 + \sigma_2^2/k)(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2}.$$

We obtain n_2 as the smallest integer greater than or equal to k times the preceding expression.

Example (sample size). Referring to {FEV.pdf}, suppose that we want 90% power for a level $\alpha = 0.05$ test of

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{against} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

subject to the constraint that $n_2/n_1 \approx 0.50$. Then

$$n_1 = 39 \approx 38.5 = \frac{(0.7234 + 0.5625/0.50)(1.960 + 1.282)^2}{0.7108^2}$$

and

$$n_2 = 20 \approx 19.2 = 0.5 \times \frac{(0.7234 + 0.5625/0.50)(1.960 + 1.282)^2}{0.7108^2}.$$

If we like, we can adjust n_1 up to 40 so that $n_2/n_1 = 0.50$ exactly.