3.

THE NORMAL DISTRIBUTION

The normal distribution is the most important distribution in psychological/statistical theory. This is due to several factors.

One factor is that many measurement characteristics in the real world have been found to be approximately normally distributed. This is true in psychology where a careful recording of observations in both observational and experimental studies has clearly defined the normal distribution.

For large sample sizes the distributions of many statistics can be approximated by the normal distribution no matter what distribution we are sampling from. A common example of this fact, that is given in undergraduate psychology courses is the Central Limit Theorem.

Another factor is that the normal distribution is mathematically tractable for much of the statistical analysis that psychologists frequently use.

Throughout the remainder of the development of this text it will be assumed that the samples are drawn from normal distributions. We will therefore require several results concerning the distribuion theory associated with sampling from the normal.

3.1 THE STANDARD NORMAL

Let z be distributed standard normal and write z ~ N(0, 1) when z has density function

$$f(z) = (2\pi)^{\frac{-1}{2}} e^{\frac{-1}{2} z^2}$$

z is distributed normally with mean 0 and variance 1.

3.2 THE GENERAL NORMAL

Suppose z ~ N(0, 1) and we put y = μ+σz where $\mu \in$ R and $\sigma \in R^+$. Then the density of y is given by

$$f(y) = (2\pi\sigma^2)^{\frac{-1}{2}} e^{\frac{-1}{2\sigma^2}(y-\mu)^2}$$

y is normally distributed with mean μ and variance σ^2, see figure 3.1. Note that the probability for the general normal can be calculated using the tables for the standard normal. In order to calculate the probability that y is less than or equal to some specified value y_o calculate

$$P(y \le y_o) = P\left(\frac{y-\mu}{\sigma} \le \frac{y_o-\mu}{\sigma}\right)$$

$$= P\left(Z \le \frac{y_o-\mu}{\sigma}\right)$$

where $Z = \frac{y - \mu}{\sigma} \sim N(0,\ 1)$

Example 1

A developmental psychologist randomly samples 50 children from a suburb of a community located beside a heavy machinery plant. The psychologist knows the norms on an auditory test for normal children; the distribution of test scores is normal with mean, μ equal to 380 and standard deviation, σ equal to 35.2. The mean score for the suburb group is 390. Test the hypothesis that the mean score of the suburb is equal to 380.

H_o: $\mu \leq 380$

H_a: $\mu > 380$

(Since the noise from the plant should increase the auditory measure indicating a decrement in capacity.)

The test statistic is $Z = \frac{\bar{y} - \mu}{\frac{\sigma}{\sqrt{n}}}$

$$= \frac{390 - 380}{\frac{35.2}{\sqrt{50}}}$$

$$= 2.01$$

Let $\alpha = .05$ for this one-tailed test and reject H_o if $Z > 1.65$. Since $2.01 > 1.65$ we reject H_o and claim that the mean of the suburb is greater than 380.

Psychologists often calculate the *observed level of significance (OLS) or p value* of a statistic. This is the probability of getting a result as far as, or farther from, the center of the distribution as that observed. In the above example Z = 2.01, we would like to find the probability of obtaining a Z statistic equal to or greater than 2.01. The p value in this example is then .0222. Typically psychological journals write p = .02 for the observed level of significance and $p < .05$ for a test of hypothesis carried out at $\alpha = .05$. Obviously the smaller the p value the greater the evidence against the null hypothesis.

A 95% confidence interval for the mean would be $\bar{y} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$,

$$= 390 \pm 1.96 \frac{35.2}{\sqrt{50}}$$

$$= (380.24, 399.76)$$

Figure 3.1

Figure 3.2

3.3 CHI-SQUARE (k) DISTRIBUTION

If Z is distributed as a standard normal variable then the distribution of Z^2 is known as a Chi-square distribution with 1 degree of freedom.

if Z ~ N(0, 1) then

$$\boxed{y = Z^2 \sim \chi^2_{(1)}}$$

An extension of the above definition is if $y_1, \ldots, y_n$ are statistically independent and $y_i \sim \chi^2_{(k_i)}$ then

$$W = \sum_{i=1}^{n} y_i \sim \chi^2_{(\sum_{i=1}^{n} k_i)}$$

In other words the distribution of the sum of a number of independent χ^2 variables is a χ^2 with degrees of freedom equal to the sum of the separate χ^2 degrees of freedom. See Figure 3.2.

Example 2

A psychologist wishes to examine a manufacturer's claim that the new polygraph has variance less than 4. From a sample of size 20 the variance s^2 was calculated as 6.2.

H_o: $\sigma^2 \leq 4$

H_a: $\sigma^2 > 4$

The alternative hypothesis indicates that the variance is actually greater than that claimed by the manufacturer.

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma^2}$$

$$= \frac{19\,(6.2)}{4}$$

$$= 29.45$$

Reject H_o if the computed χ^2 is greater than 30.1435 for $\alpha = .05$ and degrees of freedom equal to 19. The p value is about .06. It would seem then that the manufacturer's claim was substantiated. However, given the p value of .06 it would seem reasonable for the psychologist to do further testing.

3.4 STUDENT (n) DISTRIBUTION

If Z ~ N(0, 1) statistically independent of y ~ χ^2 (n) then

$$T = \frac{Z}{\sqrt{\frac{y}{n}}} \sim Student\ (n)$$

In other words, T is distributed as a t distribution. See Figure 3.3.

Example 1 (continued)

In example 1, suppose the psychologist did not know the variance, but estimated it from the sample. Let s = 30.0 from a sample of 25 children.

H_o: $\mu \leq 380$

H_a: $\mu > 380$

the test statistic is $t = \dfrac{\bar{y} - \mu}{\frac{s}{\sqrt{n}}}$

$$= \frac{390 - 380}{\frac{30}{\sqrt{25}}}$$

$$= 1.66$$

Let α = .05 for this one-tailed test and reject H_o if t(24) > 1.711. Since 1.66 < 1.711 we are unable to reject H_o.

A 95% confidence interval for the mean would be $\bar{y} \pm t(n-1)_{\frac{\alpha}{2}} \dfrac{s}{\sqrt{n}}$,

$$= 390 \pm 2.064 \ \frac{30}{\sqrt{25}}$$
$$= (377.62, 402.38)$$

3.5 F(m, n) DISTRIBUTION

If $y \sim \chi^2$ (m) statistically independent of Z $\sim \chi^2$ (n) then

$$X = \frac{\frac{(y)}{m}}{\frac{(Z)}{n}} \sim F\,(m,n)$$

See Figure 3.4 for some examples.

Example 3

Suppose a psychologist would like to compare reaction time to a visual stimulus for both men and women. The psychologist would like to see whether the variance in reaction time is the same in both groups. The variance, $s^2{}_1$, for a sample of 10 men is .105 and $s^2{}_1$ for a sample of 10 women is .058.

H_o: $\sigma^2{}_1 = \sigma^2{}_2$

H_a: $\sigma^2{}_1 > \sigma^2{}_2$

$$F = \frac{s^2{}_1}{s^2{}_2}$$
$$= \frac{.105}{.058}$$

= 1.81

Let α = .05 for a one-tailed test and reject if F is greater than F(.05, 9, 9) = 3.18. Since F is less than 3.18 we conclude that there is little evidence to reject H_o.

Figure 3.3

Figure 3.4

3.6 Exercises

1. A golf pro is interested in finding out how well the members at his course know the rules of golf. He administers a golf knowledge questionnaire to a random sample of 15 members. The scores are as follows.

   23 20 16 24 17 19 27 29 18 21
   24 19 25 26 26

   The pro knows that $\mu = 21$ and $\sigma = 2.4$ but he feels his members have more than average knowledge. Test his hypothesis.

   Construct a 95% confidence interval for the mean.

2. A clinical psychologist hypothesizes that students' scores on a depression index will tend to be higher when the workload is heavier (e.g. close to final exams). The average score on this index is 32. The psychologist administers the test to a random sample of 10 students near the end of the term. For this group, $\bar{X} = 39$ and S = 2.8. Using a one-tailed test, is the psychologist correct in his hypothesis?

   Give a 95% CI for the true mean.

3. A psychologist recently constructed a new test of reading performance. A group of 16 participants had an average score of 25, with $s_1^2 = 5.0$. The old test had $\sigma^2 = 8.0$. Does the new test have smaller variance than the old?

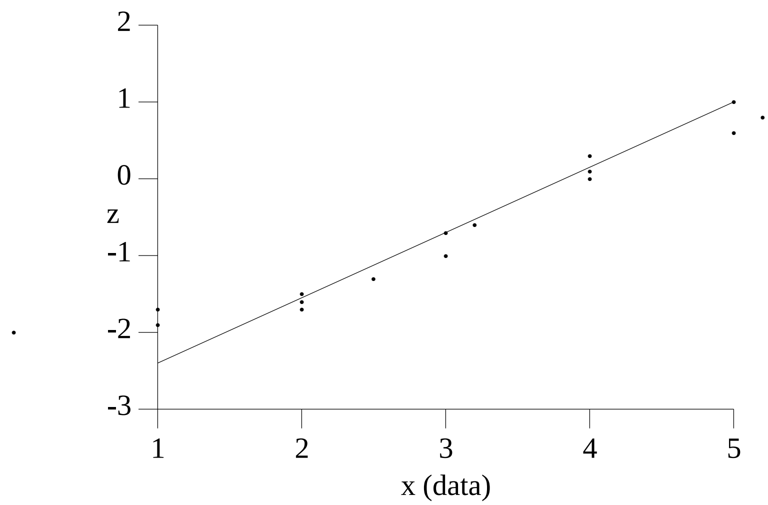
4. Suppose a second group of participants performed the test in question 3. There were 20 people in this group with $s_2^2 = 6.1$. Are the variances in both groups equal?

3.7 ASSESSING NORMALITY

A histogram can reveal non normal features of a distribution which may include skewness, groups or bimodality. Usually psychologists plot the mean, plus and minus one and two standard deviation points. This helps to compare the observed proportions in the distribution to the 68%, 95% and 99% points of the normal.

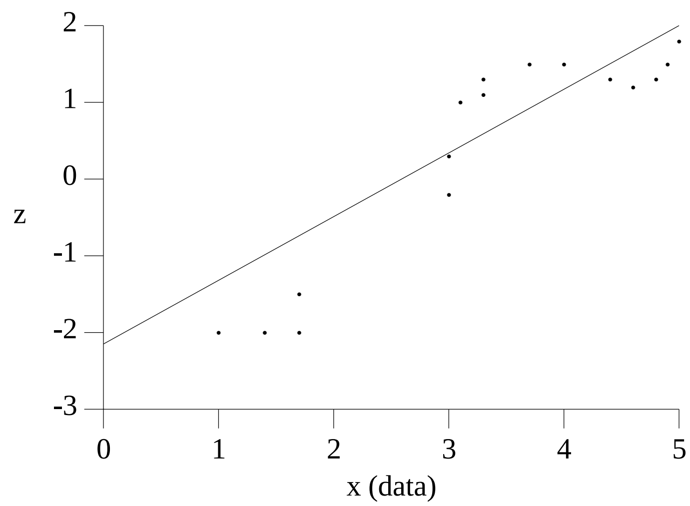
A better method of assessing normality is the *normal quantile plot* or *probability plot*. Quantile is another way of saying percentile. If we compare two similar distributions by plotting their percentiles against one another then their quantiles will be similar and the plot will be close to the line y=x. Deviations from the line reveal how the distributions differ. In the normal quantile plot each observation in the sample is seen as a quantile of the observed distribution. The smallest of 100 observations is the .01 quantile of the data because $\frac{1}{100}$ or .01 of the data lie below it. We plot the smallest observation on the x-axis and plot the corresponding z (normal) value on the y-axis. This is done for each observation. If the data follow a normal distribution then the points will fall close to the line y=x.

These plots are used by psychologists to assess normality and are usually output using computer statistical packages. Since a standard normal variable can be transformed to the general normal by a linear equation, the key feature of probability plots is that if the distribution is normal, the points will fall close to a straight line. For example,



Probability plot for normal distribution data.

The data below come from a non normal distribution. The points do not fall close to a line.



Probability plot for non normal distribution data.