Chapter Six

Empirical Rule: Let $x_1, x_2, \ldots x_n$ be a sample with mean $\bar{x}$ and standard deviation s. If the histogram is bell-shaped then:

- about 68% of the data lies in the interval $(\bar{x} - s, \bar{x} + s)$
- about 95% of the data lies in the interval $(\bar{x} - 2s, \bar{x} + 2s)$
- about 99.7% of the data lies in the interval $(\bar{x} - 3s, \bar{x} + 3s)$

Suppose you have an experiment and you repeat it n times generating specific random variable values $\{x_k\}$. Consider the following measures:

- The **sample mean**: $\bar{x} = x_1(1/n) + \ldots + x_n(1/n) = \sum_{k=1}^{n} x_k \Big/ n$.
- The **discrete variance** : $\mathrm{Var}(X) = \sum_{k=1}^{m} (x_k - \bar{x})^2 \Big/ n = \left(\sum_{k=1}^{m} {x_k}^2 / n \right) - \left(\bar{x}\right)^2$.
- The **sample variance** : $s^2 = \sum_{k=1}^{m} (x_k - \bar{x})^2 \Big/ (n-1) = \mathrm{Var}(X)\, n / (n-1)$.
- The **sample standard deviation**: $s = \sqrt{s^2}$
- The **mode**: The value of x_k with largest frequency.

Notation: Given variable X with finite mean and variance, the changed variable

$$Z = \frac{X - \mu}{\sigma}$$

yields a new data set with $\mu = 0$ and $\sigma = 1$. This new variable describes the relationship between each data value and the mean of the entire data set. Positive Z means the value lies above the mean, negative means the value lies below. The size of the z-score indicates the number of standard deviations above or below the mean.

Remark : Calculators and computer software can do these calculations for you easily and effectively.

Generalizing Measures for Grouped Data

Similar to the way we deal with discrete data, suppose you have an experiment and you repeat it n times generating random variable values $\{x_k\}$. Place these into the m classes with class marks $\{ u_1, u_2, \ldots, u_m \}$ and corresponding class frequencies $\{f_1, f_2, \ldots, f_m\}$. Then, class k has relative frequency f_k / n . Using the formulas for discrete data yields

- The sample mean: $\bar{u} = \sum_{k=1}^{m} u_k f_k / n$.

- The discrete variance: Var(X) = $\frac{\sum_{k=1}^{m}(u_k - \bar{u})^2 f_k}{n} = \frac{\sum_{k=1}^{m} u_k^{\,2} f_k}{n} - \bar{u}^2$.
- The sample variance: s_u^2 = Var(X) n / (n-1).
- The sample standard deviation: $s_u = \sqrt{s_u^2}$

Corollary: If the data is bell-shaped, then s is approximately range/6. When converting to z-scores, any data value with |Z|>3 is considered an outlier.

Group Experiment: Create several data sets from the sample space {0, 1, 2, 3, …, 9}, allowing repeats of course, whose histograms have various bell and non-bell shapes. Compute the sample mean and the standard deviation for each. Compare each to the empirical rule and see how well it applies.

Other measures of the Center and the Spread of Data:

- Median is a measure of the center. If the data were lined up, it finds the data value that would be exactly in the middle. This ignores data that lies at extreme values. Good to use if data has a few uncertain or weird values.
- Midrange = average of the largest and the smallest data is a measure of the center. This ignores data in the middle and only notices the largest and smallest values. Good to use it you are designing something where every possibility must fit.
- Range = difference of the largest and the smallest data is a measure of the spread. Easy to compute
- Inter-quartile range (IQR) = difference of the third and first quartiles is a measure of the spread. Easy to compute.

Box and Whisker Diagram: Another way to represent the distribution graphically, separating the data into quarters.

- Determine min, quartiles and max.
- Plotting these separates the data into quarters. Use boxes for the two middle quarters (box) and straight lines for the outside two quarters (whiskers).
- Note, the total width of the inside box is the IQR.
- Mark the inner fences at
    - Q1 - (1.5) IQR
    - Q3 + (1.5) IQR
- Mark the outer fences at
    - Q1 - (3) IQR
    - Q3 + (3) IQR
- To identify data that might be considered suspect:
    - Outliers are those data values that lie beyond the outer fences
    - Suspected outliers are those data values that lie between adjacent inner and outer fences.

Trimming: If the data contains outliers, one may recompute the mean and variance with the outliers removed—that is, trimming the data. In general, if one removes k values from one end of the data set, one should also remove k from the other end and then recompute the mean and variance with the smaller data set.

Winsorization: For a k-level Winsorization, to minimize the effect of extraneous values in the data set set the k smallest data values equal to x_k and and the k largest data values equal to x_{N-k+1}. Then, recompute the mean and variance with this modified data set.

Skewness: To measure whether the data is skewed to the left or right of the middle, develop a measure which is positive if the data is skewed to the right and negative is skewed to the left. Notice, when data symmetrical, generally mean = median = mode. When the data is skewed to the right, generally mean > median > mode and when data skewed to the left mean < median < mode.

- Pearson Coefficient of skewness = (mean - mode) / standard deviation. If the mode is not unique, then use 3(mean - median)/standard deviation
- Standard Measure - $\dfrac{\dfrac{\sum_{k=1}^{m}(x_k-\bar{x})^3}{n}}{s^3}$
- Bowley's Coefficient of Skewness - (Q3-Q2)(Q2-Q1)/(Q3-Q1)

Kurtosis: To measure whether the data has a sharp or flat peak, develop a measure which is positive if the data has a sharp peak and negative if the data has a flat peak.

* Standard Measure - $\dfrac{\dfrac{\sum_{k=1}^{m}(x_k-\bar{x})^4}{n}}{s^4}$

* Coefficient of Kurtosis = [(Q3-Q1)/2] / [90th percentile - 10th percentile]

Remark: Many times we have more than one set of data. We would like to compare these sets to determine if there is a significant difference between them. Assume that we have two samples which are of the same relative size--the first sample with values $x_1, x_2, \ldots x_n$ and the second sample with values $y_1, y_2, \ldots y_m$. In general, we can investigate whether the data sets are similar.

Relative Variation: $V = 100\,\sigma/\mu$. Notice, this makes distributions with large data values (and corresponding large variances) on the same level as distributions with small data values (with their correspondingly small variances.) The distribution with the small V has a greater degree of uniformity.

HOMEWORK: page 332

Questions: What if we don't know the population parameters μ, σ or p? How do we get a value that is close to them? How do we measure closeness?

Remark: We know we can take repeated samples from the distribution and obtain $\bar{x}$, s, or the relative frequency.

Defn: Valid point estimates:

- $\bar{x}$ is a good point estimate for μ
- s is a good (and *unbiased*) point estimate for σ
- the relative frequency is a good point estimate for p

Homework: page 343

Defn: A confidence interval is an open interval centered on a point estimate which has a desired probability 1-a of containing the actual parameter. We call 1 - a the confidence coefficient. For the mean □, a confidence interval will look like (xbar-a, xbar+a).

Remark: If we assume that the data is normally distributed, then the random variable xbar is also normally distributed with mean □ standard deviation $n^{1/2}/\square$. Hence, computing xbar (and if □is known) gives the confidence interval:

P(xbar-a < □ < xbar+a) = 1-a, implies
P(-a < □-xbar < a) = 1-a, implies
P(-a < xbar-□ < a) = 1-a,
P(-a $n^{1/2}/\square$ < (xbar-□) $n^{1/2}$/s < a $n^{1/2}/\square$) = 1-a, and by the normal assumption,
P(-a $n^{1/2}/\square$ < z < a $n^{1/2}/\square$) = 1-a.

Since this is now a symmetric region in N(0,1), we obtain

2 P(z < a $n^{1/2}/\square$) = 1-a, or
P(0 < z < a $n^{1/2}/\square$) = 1/2 - a/2, or
F(a $n^{1/2}/\square$) - F(0) = 1/2 - a/2, or finally
F(a $n^{1/2}/\square$) = 1 - a/2.

So, if the sample size n is given, the population standard deviation □ is given, and if the confidence level 1-a is given (and so a is known), then we can look up the z value in Table IV and find z so that F(z) = 1 - a/2. But z=a $n^{1/2}/\square$ gives the value a. Therefore, the confidence interval which contains □ is known to be (xbar-a , xbar+a).

Remark: If the data is not normally distributed to begin with, the Central Limit Theorem implies that if n=sample size is sufficiently large, the above will hold approximately and so we obtain an approximate confidence interval.

Remark: It can be shown that $n \geq 30$ is generally large enough for the CLT to hold.

Remark: We can also obtain confidence intervals for s if it is not known by using the (s-a , s+a). See section 6.5.

Section 6.7. How large should a sample be in order that xbar is certain to be a good estimate of □? We can make xbar "good" by requiring the confidence interval to have small width. So, not we let a be known and try to determine n=sample size. Notice, we can't get a value for xbar without knowing how big the sample size n is. But if we forget that we don't know what xbar is, the same calculations as above hold and give finally:

F(a $n^{1/2}/\square$) = 1 - a/2.

In the same manner as before, we can now solve for the desired n. Then, we can take the actual sample of this size, compute a real-life xbar, and see that the resulting confidence interval is indeed very close to the width we had desired it to be.

Remark: With confidence intervals, we wanted to find an interval in which we were relatively certain □ actually was in. Now, we want to guess where we think □ is and compare that with a sample value. If the sample value is very far away from what we guessed □ was, then we can conclude that our original guess was wrong. This is the essence of hypothesis testing.

Result: To test a hypothesis,
(a) Decide on the main hypothesis for □, called the null hypothesis, in the form □=some number.
(b) Decide on an alternate hypothesis for □, which will be one of □>some number, □<some number (called one-tail hypothesis) or □<> some number (called a two-tail hypothesis).
(c) Perform the same calculations as with confidence intervals except put the null hypothesis for □ in the middle. Determine the endpoint(s) of this region.
(d) Check to see where the actual sample value xbar lies. If xbar lies outside the interval above, reject the null hypothesis and accept the alternate hypothesis. Otherwise, say the test proved nothing.

Type I and Type II Errors: See paage 393

See charts on pages 411 and 412.

Remark: If n≥30, we have said that the CLT applies and we can use the normal distribution to get probabilities. What if we can not sample this many times? (eg, mummies, volcanos) If we know the original data is normally distributed, then there is no problem. However, if not, then we can use the t-distribution. See Table VI in appendix.