$$\sigma^2_{\bar{Y}} = \sum_{k=1}^{N^n} (\bar{Y}_k - \mu)^2 \Big/ N^n = \frac{\left(4(0.0)^2 + 6(\pm 0.5)^2 + 4(\pm 1.0)^2 + 2(\pm 1.5)^2\right)}{16} =$$

$$\frac{\left(4(0.0) + 6(0.25) + 4(1.0) + 2(2.25)\right)}{16} = \frac{10}{16} = 0.625$$

$\sigma = 0.7906$

Note that the histogram of the derived population shows that the population is shaped more like the normal distribution than the original population.

Probability statement from the two distributions: Find $P(1 \le Y \le 2)$

For the original, uniform population, $P(1 \le Y \le 2) = 0.5000$

For the derived population, $P(1 \le Y \le 2) = 0.6250$

THEOREM on the distribution of sample means

Given a population with mean μ and variance σ^2, if we draw all possible samples of size n (with replacement) from the population and calculate the mean, then the derived population of all possible sample means will have

Mean: $\mu_{\bar{Y}} = \mu$

Variance: $\sigma^2_{\bar{Y}} = \sigma^2 / n$

Standard error: $\sigma_{\bar{Y}} = \sqrt{\sigma^2 / n} = \sigma / \sqrt{n}$

Notice that the variance and standard deviation of the mean have "n" in the denominator. As a result, the variance of the derived population becomes smaller as the sample size increases, regardless of the value of the population variance.

Central Limit Theorem

As the sample size (n) increases, the distribution of sample means of all possible samples, of a given size from a given population, approaches a normal distribution if the variance is finite. If the base distribution is normal, then the means are normal regardless of n.

Why is this important? (It is very important!)

If we are more interested in the MEANS (and therefore the distribution of the means) than the original distribution, then normality is a more reasonable assumption.

Often, perhaps even usually, we will be more interested in characteristics of the distribution, especially the mean, than in the distributions of the individuals. Since the mean is often the statistic of interest it is useful to know that it is possibly normally distributed regardless of the parent distribution.

NOTES on the distribution of sample means

Another property of sample means

as n increases, $\sigma^2_{\bar{Y}}$ and $\sigma_{\bar{Y}}$ decrease.

$\sigma^2_{\bar{Y}} \le \sigma^2$ for any n

$\sigma^2_{\bar{Y}} < \sigma^2$ for any $n > 1$

as n increases and $\sigma_{\bar{Y}}$ becomes smaller, the distribution of the means gets closer to $\mu_{\bar{Y}}$. (i.e. we get a better estimate).

Some new terms

Reliability (as a statistical concept) – the less scatter that occurs in a set of data, the more "reliable" the estimate. This term is also associated with the word "precision" in statistics where high reliability comes from low variance and high precision. This concept is important in estimation because it suggests that data is reproducible or repeatable.

Accuracy (as a statistical concept) – this term refers to the lack of bias in the estimate, and not the smallness of the variance (e.g. reliability or precision). An accurate, or unbiased, estimated may have considerable scatter among the points, but on average the center of the distribution is neither overestimated or underestimated. This aspect of data is related to the validity of data.

Estimation of $\sigma^2_{\bar{Y}}$ and $\sigma_{\bar{Y}}$

In practice we cannot draw all possible samples.

Recall that $E(S^2) = \sigma^2$

so, $S^2_{\bar{Y}} = S^2/n$ is an estimate of $\sigma^2_{\bar{Y}}$

where;

$S^2_{\bar{Y}}$ is an estimate of the variance of sample means of size n

S^2 is the estimate of the variance of observations

$S_{\bar{Y}} = \sqrt{S^2/n} = S/\sqrt{n}$ is called the standard error to distinguish it from the standard deviation

it is also called the standard deviation of the means

Notice that this is a division by "n " for both populations and samples, not by " $n - 1$" as with the calculation of variance for samples.

$S^2_{\bar{Y}}$ is a measure of the reliability of the sample means as an estimate of the true population mean.

i.e. the smaller $S^2_{\bar{Y}}$, the more reliable $\bar{Y}$ as an estimate of μ

Ways of increasing reliability

Basically, anything that decreases our estimate of $S^2_{\bar{Y}}$ makes our estimate more reliable.

How do we decrease our estimate of $S^2_{\bar{Y}}$?

Increase the sample size; if n increases then $S^2_{\bar{Y}}$ decreases.

Decrease the variance; if our estimate of S^2 decreases then $S^2_{\bar{Y}}$ decreases.

This can sometimes be done:

- by refining our measurement techniques
- by finding a more homogeneous population to measure (stratification)
- by removing exogenous sources variance (blocking)

Notation: The variance of a variable *Y* can be denoted S^2 or S^2_Y. The subscript is only needed to clarify which variable is indicated, so it is often not needed and omitted. However, the variance of the means should be subscripted, $S^2_{\bar{Y}}$, to distinguish it from the variance of the observations.

The Z transformation for a derived population

We will use the Z transformation for two applications, individuals and means.

Applications to individual observations – to make statements about individual members of the population, such as "What percentage of the individuals would be expected to have values greater than 17?"

$$Z_i = \frac{\left(Y_i - \mu\right)}{\sigma}$$

Applications to means – to make statements about population means such as "What is the probability that the true population mean is greater than 17?"

$$Z_i = \frac{\left(\overline{Y}_i - \mu_{\bar{Y}}\right)}{\sigma_{\bar{Y}}}$$

Summary

Most testing of hypotheses will concern tests of a derived population of means.

The mean of the derived population of sample means is $\mu_{\bar{Y}}$

The Variance of the derived population of sample means is $\sigma^2_{\bar{Y}}$

The CENTRAL LIMIT THEOREM is an important aspect of hypothesis testing because it states that sample means tend to be more nearly normally distributed than the parent population. We will often work with distributions that are not normally distributed, but the fact that we are often interested in the more normally distributed means instead of the original observations allows the use of parametric statistics.

Reliability and accuracy are statistical concepts relating to variability and lack of bias, respectively.

Mean: $\mu_{\bar{Y}} = \mu$

Variance: $\sigma^2_{\bar{Y}} = \sigma^2 / n$

Standard error: $\sigma_{\bar{Y}} = \sqrt{\sigma^2 / n} = \sigma / \sqrt{n}$

Tests of hypothesis

Hypothesis – a contention based on preliminary evidence of what appears to be fact (an educated guess), which may or may not be true.

- Formulating a hypothesis is the second step in the scientific method.
- A statement of the hypothesis is the first step in experimentation.

Test of hypothesis – a comparison of the contention with a set of newly gathered data.

Hypothesis testing procedure – we will consider 7 steps

I. Set up a meaningful hypothesis such as "The population mean is equal to some value" (call it μ_0)

H_0: $\mu = \mu_0$ or $\mu - \mu_0 = 0$

This is called the null hypothesis. It is a hypothesis of equality or of no difference (even if you believe there is a difference). Note that hypotheses are always stated in terms of the population parameters, not the sample statistics we actually measure, because we are drawing inference about the population.

II. Set up an alternative hypothesis

Alternative hypotheses are denoted H_1 or H_a. This hypothesis states what is correct if the null hypothesis in not correct. This is usually the case of actual interest.

Examples:

a) H_1: $\mu \neq \mu_0$ or H_1: $\mu - \mu_0 \neq 0$ (also called the non-directional alternative)

b) H_1: $\mu < \mu_0$ or H_1: $\mu - \mu_0 < 0$

c) H_1: $\mu > \mu_0$ or H_1: $\mu - \mu_0 > 0$

III. Consider the assumptions.

1) We will be using the Z distribution, so the distribution we are testing must be normal.

2) The observations should be independent. The best guarantee of independent observations is random sampling.

3) Strictly speaking, the variance should be known in order to use the Z distribution; however it is often used for very large samples. Later we will discuss the t-distribution that is used when the variance is not known and must be estimated from the sample.

There will be a few other assumptions for other test statistics. However, the tests of hypothesis we will be using are also "robust". Statistically speaking, robustness indicates that the test performs quite well even if the assumptions are not perfectly met.

IV. Select a probability of rejecting the null hypothesis (H_0) when it is true. This is called the alpha (α) value and the value chosen is somewhat arbitrary. By convention the values usually chosen is $\alpha = 0.05$ or sometimes $\alpha = 0.01$.

for $\alpha = 0.05$ then if H_0 is true we will reject it 5% of the time, or in one of 20 samples

for $\alpha = 0.01$ then if H_0 is true we will reject it 1% of the time, or in one of 100 samples

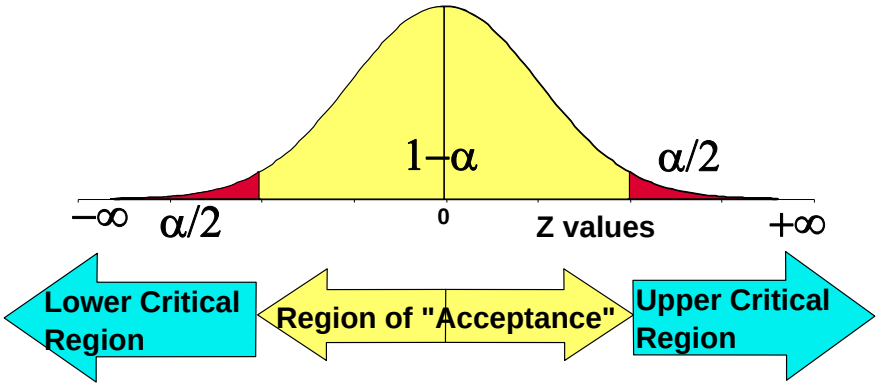
This value is sometimes called the significance level.

From this value, and the alternate hypothesis, we can determine the critical limits, those values of the test statistic that would cause us to reject the null hypothesis.

Determine a critical region, what is too large or too small by using the chosen probability or significance level.

Critical region – the area in the distribution which would lead to rejection of the null hypothesis (H_0:). When we reject we know that it is possible that the null hypothesis is true, but if it is we would only reject $\alpha*100\%$ of the time. So this type of error can be controlled.

Region of "acceptance" – the area under the distribution which would lead to "acceptance" of the null hypothesis (H_0:).



Notice that I have placed the word "acceptance" in quotes. We cannot really state that we "accept" the null hypothesis because it is also possible that we would be wrong in doing so. Unfortunately, in practice the probability of this type of error is unknown and therefore one cannot "accept" with a known probability of error (more later under Type II error and Power).

V. Draw a sample from the population of interest (as defined by the investigator), and

a) Compute an estimate of the parameter in the hypothesis; in our example the hypothesis was about μ so the statistic will be $\bar{Y}$, recall $E(\bar{Y}) = \mu$.

b) The value of $\bar{Y}$ from the sample now becomes one of many possible observations from the derived population of all sample means.

c) Recall that the derived population has,

$$\mu_{\bar{Y}} = \mu$$

$$\sigma^2_{\bar{Y}} = \sigma^2 / n$$

$$\sigma_{\bar{Y}} = \sqrt{\sigma^2 / n} = \sigma / \sqrt{n}$$

d) Recall that the distribution of sample means ($\bar{Y}_k$) approaches a normal distribution as the value of n increases (according to the Central Limit Theorem). This helps meet our assumption of normality.