# Introduction to Applied Statistics

<https://newonlinecourses.science.psu.edu/stat800/node/1/>

<https://newonlinecourses.science.psu.edu/stat414/node/318/>

## Designing samples

Four common probability sampling schemes are:

1. **Simple Random Sampling (SRS): sample without replacement**
2. **Stratified Random Sampling**

The population is divided into important subgroups (e.g. by country; by gender) which are groups of individuals or subjects that are similar in a way that may affect their response.

Then separate simple random samples are taken from each subgroup. These subgroups are called **strata**. This is done to be sure every important subgroup is represented properly in the overall sample which will enhance the efficiency of this design.

1. **Cluster Sampling**

The population is divided into several subgroups by geographic proximity or closeness of individuals to each other on a list. These subgroups are called **clusters**. Then some clusters are randomly picked to be in the sample. There may be further random sampling of individuals within the selected clusters. For instance, for an on-campus survey, we might randomly pick a few dorms and only include some or all of the students from those dorms in the survey.

Cluster sampling differs from Stratified sampling in that:

* Cluster sampling is not initially concerned with similarities among the individuals. However, once the clusters are created one may have to account for any possible similarities.
* In stratified sampling we create the subgroups based on some criteria – ethnicity, geographic region, and then random sampling of individuals or subjects is done. In Cluster sampling, clusters of individuals or subjects are randomly sampled.



1. **Multistage Sampling**

Selects successively smaller groups from the population, ending with clusters of individuals. Most opinion polls are done in stages. For example, you may start by splitting your home state into regions. Then stratify within each region by rural, suburban, and urban. From these strata you would randomly select some communities, from which these would be divided by some fixed area (think by city blocks) – i.e. clusters. Finally, within these clusters all individuals would then be sampled.

## Basic principles of statistical design of experiments

1. In a good design, we **control** for the effects of outside variables. To avoid confounding the effect of the treatment with other variables, a comparison should be made.
2. **Randomization** – relying on chance to decide which subjects are studied **(random selection)** and which get what treatment **(random assignment)** – is usually the best way to ensure that bias does not creep into the initial selection.

The random assignment is required for an experiment.  When both random assignment and selection are part of the study then we have a completely randomized experiment.

Without random assignment (i.e.an observational study) then the treatment can only be referred to as being **related** to the outcome, and we can’t conclude as **causal effect**.

1. **Replication** is essential: each treatment should be repeated on a large enough number of units to allow systematic effects to be seen. We need to distinguish between outcomes that are statistically different or just different due to random chance or sampling noise.

## Types of bias

1. **Non-response** – large percentage of those sampled do not to respond or participate.
2. **Response**– when study participants either do not respond truthfully or give answers they feel the researcher wants to hear.  For example, when students are asked if they ever cheated on an exam even those who have would respond with "no".
3. **Selection**–  this bias occurs when the sample selected does not reflect the population of interest.  For instance, you are interested in the attitude of female students regarding campus safety but when sampling you also include males.  In this case your population of interest was female students however your sample included subject not in that population (i.e. males).

## Discrete variables

* Probability distribution: probability of certain value.
* Cumulative probability:
* Expected value (Mean): the phrase **expected value**is a synonym for **mean** value in the long run (meaning for many repeats or a large sample size).
* Standard Deviation (variance)

## Binomial random variable

* Probability is:
* Expected value and standard deviation for # of success is:

## Continuous random variable

To describe probabilities for a continuous random variable, we use a probability density function. A **probability density function**is a curve such that the area under the curve within any interval of values along the horizontal gives the probability for that interval.

Normal distribution has an empirical 68-95-99.7 rule.

The cumulative probability for a value equals the cumulative probability for that value's z-score.

## Normal approximation to the binomial

This approximation can take place as long as:

1. The population size must be **at least**10 times the sample size.
2. **and**. [These constraints take care of population shapes that are unbalanced because *p*is too close to 0 or to 1.]

## Population parameters vs. sample statistics

The statistic is used to estimate the parameter. The statistic can vary from sample to sample, but the parameter is understood to be fixed.

The statistic, then, can take on various values depending on the result of repeated random sampling. The distribution of these possible values is known as the **sampling distribution.**

|  |  |  |
| --- | --- | --- |
| **Parameter Names and Description** | **Symbol for the Population Parameter** | **Symbol for the Sample Statistic** |
| **For Categorical Variables:**  One population proportion (or probability) | *p* |  |
| Difference in two population proportions | *p*1 - *p*2 |  |
| **For Quantitative Variables:**  One population mean | μ |  |
| Population mean of paired differences (dependent or paired) | μ*d* |  |
| Difference in two population means (independent) | μ1 - μ2 |  |

* Paired or dependent: the observations are taken on the same individual or two similar individuals (same population).
* Independent: taking observations from two distinct groups (two populations).

## Sampling distribution of sample statistics

Sample statistics are random variables and therefore vary from sample to sample. As a result, sample statistics also have a distribution called the **sampling distribution**. These sampling distributions, similar to distributions discussed previously, have a mean and standard deviation. However, we refer to the standard deviation of a sampling distribution as the **standard error**.

The standard error is simply the standard deviation of a sampling distribution.

## Sampling distributions for sample proportion

If numerous repetitions of samples are taken, the distribution of  is said to approximate a normal curve distribution.

In general, if  and , the sampling distribution of  is about normal with mean of p and standard error .

For example, Suppose the proportion of all college students who have used marijuana in the past 6 months is . For a class of size , representative of all college students on use of marijuana, what is the chance that the proportion of students who have used mj in the past 6 months is less than .32 (or 32%)?  () ⬄ Given the population proportion is 0.4, what is the chance to observe a sample proportion of 0.32 for 200 samples?

**Solution**:

The sample proportion follows a normal distribution with

Therefore,

Which is very unlikely.

If our hypothesis is the population proportion is 0.4, based on our observation, we will reject the null hypothesis and we won’t believe a claim that 40% of college students used mj in the past 6 months. The proportion should be less than 0.4.

## Sampling Distribution for the sample mean

The **central limit theorem**states that if a large enough sample is taken (typically ) then the sampling distribution of  is approximately a normal distribution with a mean of and a standard deviation of . Since in practice we usually do not know or σ we estimate these by  and  respectively. In this case  is the estimate of and is the standard deviation of the sample. The expression   is known as the standard error of the mean, labeled .

**i. The Law of Large Numbers**says that as the sample size increases the sample mean will approach the population mean.

**ii. The Central Limit Theorem**says that as the sample size increases the sampling distribution of  approaches the normal distribution. We see this effect here for . Generally, we assume that a sample size of  is sufficient to get an approximate normal distribution for the distribution of the sample mean.

**iii. The Central Limit Theorem**is important because it enables us to calculate probabilities about sample means.

## Comparing two groups

Responses in each group must be **independent** of those in the other, meaning that different, unrelated, unpaired individuals make up the two samples. Sample sizes, however, may vary between the two groups.

To look at the difference between the two groups, we look at the difference between population parameters for the two populations involved.

* For categorical data, we compare the proportions with a characteristic of interest (say, proportions successfully treated with two different treatments).
* For quantitative data, we compare means (say, mean GPAs for males and females).

When comparing two means for independent samples, our initial thought goes to how do we calculate the standard error. The answer depends on whether we can consider the variances (and therefore the standard deviations) from each of the samples to be equal (**pooled**) or unequal (**unpooled**). This implies that prior to doing a two-sample test we will need to first find the standard deviation for each sample.

RULE OF THUMB - If the larger standard deviation is **no more than twice** the smaller standard deviation, then we would consider the two population variances equal.

The following summarizes population and sample notation for comparisons, and gives the null hypothesis for each situation (proportions and means).

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter name and description** | **Symbol for population parameter** | **Typical null hypothesis** | **Symbol for the sample statistic** |
| **Categorical** Response Variable Difference in two population proportions | *p*1 – *p*2 | *H*0: *p*1 – *p*2 = 0 (or some value) |  |
| **Quantitative** Response Variable Difference in two population means | μ1− μ2 | *H*0: μ1− μ2 = 0 (or some value) |  |
| **Quantitative** Response Variable Difference between matched pairs | μd | *H*0: μd= 0 (or some value) |  |

Remember that hypotheses are statements about populations!

## Types of experiment design & tests

1. **Independent measures / between groups (2 sample t-test)**: Different participants are used in each condition of the independent variable.

t-test for mean variances from control and treatment groups, respectively. (2-sample t-test)

If the sample sizes and variances are the same for both groups,

There are two options for estimating the variances for the independent samples (unequal sample size in each group):

* Using pooled variances

The pooled variance (equal variance) is:

* Using unpooled (or unequal) variances

The unpooled variance (unequal variance) is:

When to use which? When we are reasonably sure that the two populations have nearly equal variances, then we use the pooled variances test. Otherwise, we use the unpooled (or separate) variance test.

t statistic follows a t distribution with significance at degree of freedom.

1. **Repeated measures /within groups**: The same participants take part in each condition of the independent variable.
2. **Matched pairs**: Each condition uses different participants, but they are matched in terms of important characteristics, e.g., gender, age, intelligence, etc.

t-test for the mean difference from each matched pair and see if it’s significantly different than zero.

A paired t-test is just a 1-sample t-test on the paired differences.

## Procedure for carrying out a paired t-test

The null hypothesis is that the difference is zero.

The alternative hypothesis is that the difference is not zero

1. Calculate the difference (di = yi − xi) between the two observations on each pair, making sure you distinguish between positive and negative differences.
2. Calculate the mean difference, .
3. Calculate the standard deviation of the differences, , and use this to calculate the standard error of the mean difference,
4. Calculate the t-statistic, which is given by

Under the null hypothesis, this statistic follows a t-distribution with n − 1 degrees of freedom.

1. Use tables of the t-distribution to compare your value for T to the distribution. This will give the p-value for the paired t-test.
2. Confidence interval for the true mean difference

Where,

We can be 95% sure that the true mean lies somewhere between our CI.