# Sample Statistics

Suppose we have a (finite) population with a characteristic having values $x_1, x_2, \cdots, x_N$. Here $x_i$ is the value of the characteristic for the $i$-th member of the population. Consider for example, a population consisting of just three trees, A, B, and C and suppose that the characteristic of interest is the weight of the tree. Suppose the weight of tree A is $x_1 = 9$, of B is $x_2 = 15$ and of C is $x_3 = 9$.
Population parameters are computed as usual:
**Population mean parameter:**

$$\mu = \frac{\sum_i x_i}{N}$$

**Population deviation parameter:**

$$\sigma = \sqrt{\frac{\sum_i (x_i^2)}{N} - \mu^2}$$

**Population proportion parameter:**

$$\theta = \frac{\text{number of successes}}{N}$$

where in the latter case the population is split into just two categories by the characteristic, one called *success* and the other *failure*.
Other useful parameters are those for median ($M$), mode ($MO$) and variance ($V$). In the example, $\mu = (9+15+9)/3 = 11$, $\sigma = \sqrt{(9^2+15^2+9^2)/3 - 11^2} \approx 2.83$ and $\theta = 1/3$ if we call values $\geq 10$ successes and the others failures, thus tree B produces the only success.
The goal is to estimate these parameters via a sample. Select a sample of $n$ elements. We consider just two possible methods. In both the probability function for the samples selected is uniform; when the population is finite this means all possible samples are equally likely. In **independent random sampling (irs)**, the sampling is done with replacement, while in **simple random sampling (srs)** the sampling is done without replacement. It is easier to analyze the case of an independent random sample since the successive selections are independent; however, often sampling is done without replacement.
Let $X_1, X_2, \cdots, X_n$ be the random variables that keep track of which characteristic values where selected: $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$. For example, if we selected a sample of size 2 which consisted of B and C then $x_{(1)} = 15$ and $x_{(2)} = 9$. This example is simple minded because in practice both $n$ and $N$ are larger with $N$ being fairly large, since if $N$ is not large a census is likely feasible.
A **sample statistic** is a random variable that is a function of the sample random variables $X_i$. *The value of a "good"† sample statistic on a specific sample is used to estimate a population parameter.* Here are three important statistics:
**The sample mean statistic** is $\overline{X} = \sum_i X_i/n$, thus on a sample $x_{(1)}, \cdots, x_{(n)}$ it has value equal to the ordinary average of the sample values:

$$\overline{x} = \frac{\sum_i x_{(i)}}{n}$$

† Some desirable properties for a statistic are that it be unbiased, minimum variance, consistent and sufficient (see page 3).

**The sample deviation statistic** is $S = \mathrm{CF} \cdot \sqrt{\frac{n}{n-1}}\sqrt{\frac{\sum_i (X_i^2)}{n} - \overline{X}^2}$, thus on a sample $x_{(1)}, \cdots, x_{(n)}$ it has value

$$s = \mathrm{CF} \cdot \sqrt{\frac{n}{n-1}}\sqrt{\frac{\sum_i (x_{(i)}^2)}{n} - \overline{x}^2}.$$

This is the ordinary deviation of the sample *corrected by the factor* $\sqrt{\frac{n}{n-1}}$, a factor which compensates for the "clipping" inherent in the sample relative to the population. An additional Correction Factor, $\mathrm{CF} = \sqrt{(N-1)/N}$, appears only if the sampling is done without replacement and the population $N$ is finite, but even then it is negligible ($\geq .975$) and so often ignored if $N \geq 30$. Note that for $n = 1$, $S$ is undefined. This is no surprise because you can not realistically guess a deviation if you only have one sample value.
**The sample deviation statistic** is $P =$ (the number of successes in the $X_i)/n$, thus on a sample it has value $p =$ (the number of successes among the $x_i)/n$.

These statistics are good statistics for estimating the corresponding population parameters. *The best estimate† for $\mu$, $\sigma$ and $\theta$ are the respective values $\overline{x}$, $s$ and $p$ of the sample statistics $\overline{X}$, $S$ and $P$ on the sample $x_{(1)}, \cdots, x_{(n)}$.* In the example, based on the sample $x_{(1)} = 15$ and $x_{(2)} = 9$ the best estimate for $\mu$ is $\overline{x} = (15+9)/2 = 12$, the best estimate for $\sigma$ is $s = \sqrt{2/3}\sqrt{2/1}\sqrt{(15^2+9^2)/2 - 12^2} = \sqrt{2/3}\sqrt{2}\sqrt{9} \approx 3.46$, and the best estimate for $\theta$ is $p = 1/2$.
Actually the proportion and the mean are closely related. Define $x^*$ to be 1 if $x$ is a success and 0 otherwise. Then the proportion parameter and statistic for $x$ are just the mean parameter and statistic, respectively, for $x^*$. That is, $\theta = \mu$, $P = \overline{X^*}$ and $p = \overline{x^*}$.
*$\overline{X}$, $S$ and $P$ are random variables so they have distributions.* These are distributions of the estimates — different samples give different estimates and each estimate has a certain probability of being the estimate you will give. For instance, consider simple random samples of size 2 in the example.††

| Sample | $x_{(1)}$ | $x_{(2)}$ | $\overline{x}$ | $s$ | $p$ |
|---|---|---|---|---|---|
| A,B | 9 | 15 | 12 | $\approx 3.46$ | .5 |
| A,C | 9 | 9 | 9 | 0 | 0 |
| B,C | 15 | 9 | 12 | $\approx 3.46$ | .5 |

Each sample is as likely as any other since the sample is a simple random sample, thus $\overline{X}$ has distribution:

| $\overline{X}$ | 9 | 12 |
|---|---|---|
| $P(\overline{X})$ | 1/3 | 2/3 |

Random variables have averages and deviations:

$$\mu_{\overline{X}} = 9 \cdot 1/3 + 12 \cdot 2/3 = 11$$

† $\overline{X}$ and $P$ are unbiased, minimum variance, consistent and sufficient for $\mu$ and $\theta$ as is $S^2$ for $\sigma^2$. $S$ is somewhat biased for $\sigma$, but it is used anyway.

†† Notice when replacement is allowed, that there are 9 possible samples. A sample AA occurs only once, while a "mixed" sample can occur twice: AB or BA. Accordingly the distribution for $\overline{X}$ is 9, 12, 15 with probabilities 4/9, 4/9, 1/9.

$\mu_{\overline{X}}$ is the **average estimate** for the population parameter $\mu$.

$$\sigma_{\overline{X}} = \sqrt{9^2 \cdot 1/3 + 12^2 \cdot 2/3 - 11^2} \approx 1.41$$

$\sigma_{\overline{X}}$ is the spread or **deviation in estimates** for the population parameter $\mu$, so it is a measure of how good the estimate $\overline{x}$ is for $\mu$.
It is important to keep in mind that in a sampling situation such as above, you will not actually explicitly know $\overline{X}$, $\mu_{\overline{X}}$, $\sigma_{\overline{X}}$, $X$, etc. unless additional assumptions are made. All you know is the sample random variable values $x_{(i)}$.
Similarly

| | | |
|---|---|---|
| $S$ | $0$ | $\approx 3.46$ |
| $P(S)$ | $1/3$ | $2/3$ |

| | | |
|---|---|---|
| $S^2$ | $0$ | $12$ |
| $P(S^2)$ | $1/3$ | $2/3$ |

| | | |
|---|---|---|
| $P$ | $0$ | $1/2$ |
| $P(P)$ | $1/3$ | $2/3$ |

The average estimate for population deviation $\sigma$ is $\mu_S = 0 \cdot 1/3 + 3.46 \cdot 2/3 \approx 2.31$.
This value has a bias relative to the true value of about 2.83. On the other hand, $\mu_{S^2} = 8$ which is also $\sigma^2$ so there is no bias there.
The deviation in the estimate for population deviation $\sigma$ is
$\sigma_S = \sqrt{0^2 \cdot 1/3 + 3.45^2 \cdot 2/3 - \mu_S{}^2} \approx 1.63$.
The average estimate for population proportion $\theta$ is $\mu_P = 0 \cdot 1/3 + 1/2 \cdot 2/3 = 1/3$.
The deviation in the estimate for population proportion $\theta$ is
$\sigma_P = \sqrt{0^2 \cdot 1/3 + (1/2)^2 \cdot 2/3 - \mu_P{}^2} \approx .236$.

PROPERTIES OF A GOOD STATISTIC:
A statistic $Y$ is said to be **unbiased** for a population parameter $\alpha$ if the average estimate $\mu_Y$ is the true value $\alpha$. It is **minimum variance** if $\sigma_Y$ is as small as possible. It is **sufficient** if, roughly, knowing the actual sample values reveals no more about $\alpha$ than merely knowing the value of the statistic $Y$. It is **consistent** if, roughly, as the sample size $n$ gets larger the value of the statistic approaches $\alpha$.

THEOREM
Suppose a population has mean parameter $\mu$ and deviation parameter $\sigma$ and suppose a simple or independent random sample of size $n$ is selected. Then the mean statistic $\overline{X} = (X_1 + \cdots + X_n)/n$ satisfies $\mu_{\overline{X}} = \mu$ and $\sigma_{\overline{X}} = \text{SPCF} \cdot \sigma/\sqrt{n}$ where the Small Population Correction Factor, $\text{SPCF} = \sqrt{(N-n)/(N-1)}$, appears only if the sampling is done without replacement and the population $N$ is finite, but even then it is negligible ($\geq .975$) if $n \leq 5\%N + 1$.
CENTRAL LIMIT THEOREM
In addition, $\overline{X}$ is appoximately normally distributed with mean $\mu_{\overline{X}} = \mu$ and deviation $\sigma_{\overline{X}}$.

REMARKS: The approximation gets better the larger $n$ is, but as a general rule of thumb, it is good enough if $n \geq 30$.
If the population is normally distributed then $\overline{X}$ is exactly normally distibuted.
Note that as $n$ gets larger, the deviation $\sigma_{\overline{X}}$ approaches zero.