NRES 798 — Statistical Methods for Ecologists
Chapter 3: Summary Statistics

Oscar García

February 11, 2013

Contents

1 Samples and statistics

A *population*, or population model, is a probability distributions. It usually involves one or more unknown parameters.

Given a population with random variable Y, a (simple) random *sample* is a list or vector of n RVs (*observations*)

$$(Y_1, Y_2, \ldots, Y_n) \ ,$$

where the Y_i are independent, and Y_i has the same distribution as Y [1].

A *statistic* is any function of the sample that does not depend on the parameters. It follows that a statistic is an RV, and its distribution can be derived from the distribution of the population. As is customary in math, sometimes the same word is used both for the function and for the function values (numbers). Which is which should (hopefully) be clear from context.

Statistics can be used for purely descriptive purposes, as data summaries, and/or for statistical inference. An *estimator* is a statistic used to compute parameter estimates.

The standard deviation of an estimator is called its *standard error*.

2 The empirical distribution

A trick that is sometimes useful is to define the *empirical distribution* for a sample. It is simply a discrete distribution that gives equal probabilities to the observed Y_i's. That is,

$$f_e(y) = \frac{1}{n} \quad \text{for } y = Y_1, Y_2, \ldots, Y_n \ .$$

Of course, this PDF looks nothing like the PDF of Y, at least not if Y is continuous. But the CDF $F_e(y)$, a step function that jumps by $1/n$ at each of the observations, approximates the CDF of Y as n increases (Figure 1 [2]).

[1] For some reason Y is used instead of X in this Chapter.

[2] Actually, this shows a particularly good looking sample, most samples of 20 look much worse (try it!). More on this later.

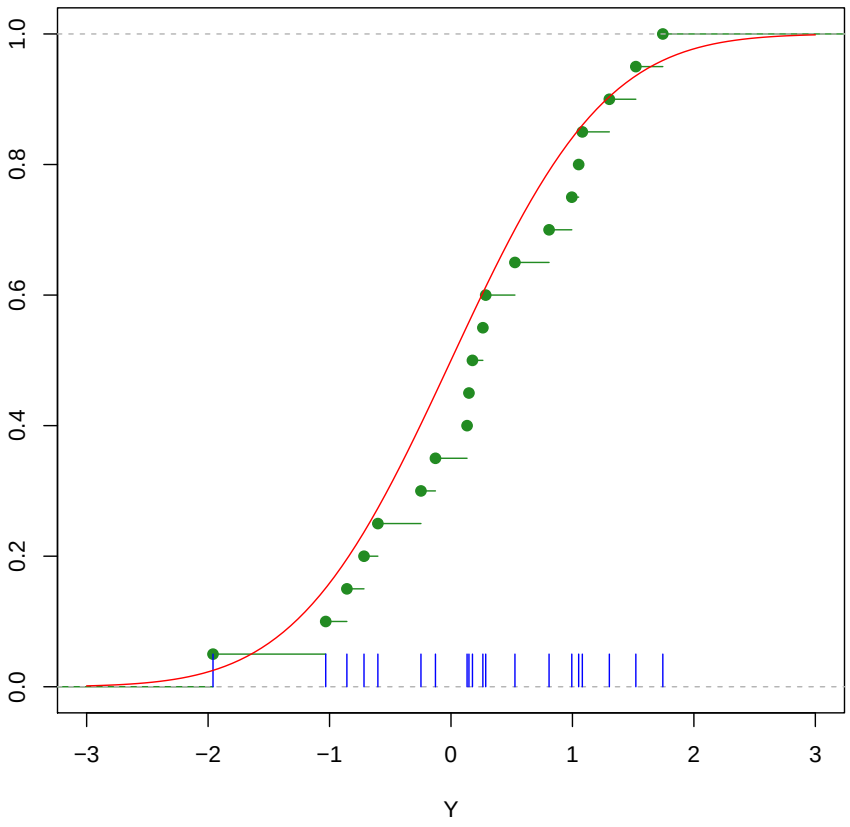


Figure 1: Empirical PDF and CDF for a random sample of $n = 20$ values from a normal distribution, plotted together with the normal CDF.

It should be clear that $F(y)$ is the proportion of observations in the sample that are less than or equal to y.

It may be convenient to consider the observations as sorted in ascending order of size. This *empirical CDF*, or *empirical distribution function* (EDF), has a number of uses, among other dealing with medians and sample quantiles later in this chapter. It can be calculated in *R* with the function `ecdf`.

3 Location

3.1 Arithmetic mean

Consider a (simple random) sample of n independent observations Y_i from a population with mean $E[Y] = \mu$ and variance $V[Y] = \sigma$. The sample *arithmetic mean*, or simply *mean*, is

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \ .$$

From the linearity of the expectation,

$$E[\overline{Y}] = \frac{1}{n} \sum_{i=1}^{n} E[Y_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu \ .$$

$\overline{Y}$ is an *unbiased* estimator of μ.

From the properties of the variance (and the independence of the Y_i),

$$V[\overline{Y}] = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} V[Y_i] = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n} \ .$$

It is seen that as n increases, the sample variance decreases. The sample mean tends to the population mean μ. This is a form of the *Law of Large Numbers* [3].

From the previous equation, the standard error of the mean, that is, its standard deviation, is

$$s_{\overline{Y}} = \sqrt{V[\overline{Y}]} = \frac{\sigma}{\sqrt{n}} \ .$$

[3] We have not used the assumption of normality from the textbook. It is only necessary that μ and σ^2 are finite. This is usually the case, except in pathological examples such as the Cauchy distribution.

From the Central Limit Theorem, we also know that the distribution of $\overline{Y}$ tends to a normal $N(\mu, \sigma/\sqrt{n})$ as n becomes large.

Note that the sample mean is the expected value of the empirical distribution:

$$\sum Y_i f_e(Y_i) = \sum Y_i \frac{1}{n} = \overline{Y} \; .$$

3.2 Geometric mean

It could be argued that if an animal population grows one year at 10%, and the next year at 20%, the "proper" two-year average growth should not be the arithmetic mean 15%. Why? If the initial population size[4] is N, after the first year it is $k_1 N$, where the growth factor is $k_1 = 1 + 10/100 = 1.1$. After the second year the population size is $k_2 k_1 N$, with $k_2 = 1 + 20/100 = 1.2$. A more "meaningful" summary for the growth over the two years might be a growth factor $\hat{k}$ such that $k_2 k_1 N = \hat{k}\hat{k}N$, that is, $\hat{k} = \sqrt{k_1 k_2}$.

As the example suggests, in some situations, typically involving multiplicative processes, a location statistic using multiplication instead of addition may be more useful, the *geometric mean*

$$\mathrm{GM} = \sqrt[n]{\prod_{i=1}^{n} Y_i} = \left(\prod_{i=1}^{n} Y_i \right)^{\frac{1}{n}} \; .$$

Note that, taking logarithms,

$$\log \mathrm{GM} = \frac{1}{n} \sum_{i=1}^{n} \log Y_i = \overline{\log Y_i} \; ,$$

so that

$$\mathrm{GM} = \exp(\overline{\log Y}) \; .$$

3.3 Harmonic mean

Sometimes the *harmonic mean* can be useful:

$$\mathrm{H} = \frac{1}{\overline{1/Y}} \; .$$

[4] Do not confuse with the statistical population!

It can be shown that always

$$\mathrm{H} \leq \mathrm{GM} \leq \overline{Y} \, .$$

3.4 Transformations

The arithmetic mean tends to work well when the distribution is close to symmetric. It may be possible to transform the original variables through some function g, so that $g(Y_i)$ is more symmetric than Y_i. One can then use the mean of the transformed variables to obtained a transformed mean, say M, such that

$$\begin{aligned} g(\mathrm{M}) &= \overline{g(Y)} \\ \mathrm{M} &= g^{-1}[\overline{g(Y)}] \end{aligned}$$

The geometric and harmonic means are special cases of this, with $g(Y) = \log(Y)$ and $g(Y) = 1/Y$, respectively. Presumably they would work well when those transformations help to make the distribution more symmetric.

3.5 Median, mode

The sample *median* is the mid-point of the data, a value that has an equal number of observations below and above it. It corresponds to the point in the middle if n is odd. If n is even, it can be said that any value between the two central points is a median, or more often, it is defined somewhat arbitrarily as the mean of those two points

More generally, the median of any RV, discrete or continuous, is the value that divides the probability in half. In other words, the value y_m for which the CDF $F(y_m) = 0.5$. Or $y_m = F^{-1}(0.5)$. The sample median is the median of the empirical distribution (see Figure 1).

The median may be more meaningful than the mean in describing the location of the "centre" of a skewed distribution. For this reason it is often used in official statistics for things like median income, or median house prices. Note that it is invariant under transformations, that is, the median of $g(Y)$ is the g of the median of Y. Except for a small difference with even n, due to the averaging of the 2 central values.

Another advantage is the resistance to outliers and observation errors. A gross error in one observation either does not change the median, or it changes it only slightly if the correct value was on the opposite side.

A disadvantage compared to the mean is that the mean is much more convenient mathematically. Derived distributions are simpler for the mean than for the median. In addition, the median is usually more variable; for the normal distribution the standard deviation of the sample median is about 25% larger than that of the mean.

For RVs, the mode is the value for which the PDF has a maximum. In the same way, for discrete populations the *sample mode* is the most frequent value in the sample. In principle, samples from a continuous population do not have a mode, since values do not repeat (note also that the empirical PDF is flat). Sometimes a mode is given based on rounding, or on binning in a histogram.

4 Spread, variability

4.1 Variance, standard deviation

The sample variance is variously defined as the sum of the squared deviations from the mean, divided by either n or by $n-1$. There is endless confusion about which denominator to use. This is compounded with the mystical powers of the concept of *degrees of freedom* [5].

The textbook adopts the terminology *mean square* for

$$s^2 = \frac{1}{n}\sum (Y_i - \overline{Y})^2 \ ,$$

and *sample variance* for

$$s^2 = \frac{1}{n-1}\sum (Y_i - \overline{Y})^2$$

[5] The name *degrees of freedom* comes from some obscure physics analogy. An object free to move in 3 dimensions is said to have 3 degrees of freedom. If it moves on a table, there is one constraint, fixed z, and is free to move along x and y: 2 degrees of freedom. If it is on rails, there are two constraints, and $3-2=1$ degree of freedom. In computing the sample variance, the mean is one constraint: fixing it, the observations cannot take on any arbitrary values, they have to sum to n times the mean; they have $n-1$ "degrees of freedom". In other instances one estimates a certain number of parameters, and it is then said that the number of degrees of freedom is n minus the number of parameters.

(but uses the same symbol). We will use this convention, just be aware that this is not standardized.

Anyway, note that the mean square is the variance of the empirical distribution, by the general definition of variance of RVs.

One justification for the use of $n-1$ in the sample variance is that then $E[s^2] = V[Y] = \sigma^2$, that is, it is an unbiased estimator of the population variance [6]. On the other hand, the standard error of the mean square is lower. In fact, it can be shown that using a denominator $n+1$ gives a standard error even lower, at the cost of more bias. It seems common to believe that any bias is "bad", but a recurring theme throughout statistical theory is the balancing of bias and variance.

The *sample standard deviation* is the square root of the sample variance. The square root of the mean square is frequently called the *root mean square* (RMS). Both are biased estimators of the population standard deviation [7].

4.2 Standard error of the mean

As shown in Section 3.1, the standard error (SE) of the mean (its standard deviation) is $\sigma/\sqrt{n}$. Because σ is generally unknown, the SE is estimated by

$$s_{\overline{Y}} = \frac{s}{\sqrt{n}} \,,$$

which is also called the *standard error of the mean*, or more precisely, the estimated standard error.

There should be no confusion with the standard deviation (SD) s, they are different things. The SD s is an estimate of σ, the SD of the population or, equivalently, the SD of a single observation Y_i. The SE $s_{\overline{Y}}$ is an estimate of the SD of the mean $\overline{Y}$, which is obviously smaller.

---

[6] $E[\sum(Y_i - \overline{Y})^2] = \sum E[((Y_i - \mu) - (\overline{Y} - \mu))^2] = \sum E[(Y_i - \mu)^2 - 2(Y_i - \mu)(\overline{Y} - \mu) + (\overline{Y} - \mu)^2] = \sum(\sigma^2 - 2\sigma^2/n + \sigma^2/n) = n\sigma^2(1 - 1/n) = \sigma^2(n-1)$. The middle term in the binomial expansion works out like this: $2E[(Y_i - \mu)(\overline{Y} - \mu)] = 2E(Y_i - \mu)\sum_j (Y_j - \mu)/n] = 2E[(Y_i - \mu)^2]/n = 2\sigma^2/n$, because $E[(Y_i - \mu)(Y_j - \mu)] = \mathrm{Cov}(Y_i, Y_j) = 0$ for $i \neq j$.

[7] For any function g, $E[g(Y)] \neq g(E[Y])$, unless g is linear or the variance is 0. *Jensen's inequality* is more specific: if g is convex (the curve bulges downward), then $E[g(Y)] \geq g(E[Y])$, if g is concave, then $E[g(Y)] \leq g(E[Y])$. The square root is concave, therefore $E[s] < \sigma$.

4.3 Moments, skewness, kurtosis

For an RV, the r-th central *moment* is the expected r-th power of the deviations from the mean. For a sample, it is the r-th central moment of its empirical distribution:

$$\mathrm{CM}_r = \frac{1}{n}\sum (Y_i - \overline{Y})^r \ .$$

Clearly, $\mathrm{CM}_1 = 0$, and CM_2 equals the mean square, or the variance in the case of the RV or population.

The *skewness* (for both samples and populations) is $g_1 = \mathrm{CM}_3/\mathrm{CM}_2^{3/2}$, and is a measure of asymmetry. For a normal distribution $g_1 = 0$.

The *kurtosis* is $g_2 = \mathrm{CM}_4/\mathrm{CM}_2^2 - 3$, where the 3 is subtracted so that for a normal $g_2 = 0$. It is said that it measures how flat the PDF is, or how heavy the tails are, although there has been some discussion in the literature on if this is always true.

As seen in the Lab, the variabilities of g_1, and especially g_2, are rather large, so that they may not be all that useful in practice[8].

4.4 Quantiles

Quantiles are RV values that correspond to a given probability in the CDF. That is, a p-quantile is $F^{-1}(p)$. Same for samples, using the empirical CDF. Various conventions are used in samples and discrete distributions for values "in-between", see the *R Help* for `quantil` (remember what happens with the median if n is even).

They can be used to summarize spread (and also location). Special cases are the median ($p = 0.5$), *quartiles* ($p = 0.25, 0, 5, 0.75$), *deciles* ($p = 0.1, 0.2, \ldots, 0.9$), the minimum ($p = 0$), and the maximum ($p = 1$) [9].

[8] Moments higher than the second or third, and more generally, the determination of distribution shape, require samples of astronomical size to get a reasonable precision. The variability tends to be underestimated by many researchers, who put too much faith on distributions obtained by sampling. For instance, tree diameter distributions in forest ecology. In fact, I cheated a bit going through several samples to get a decent picture for Figure 1; if you try it you will see that most samples look rather horrible.

[9] Common measures of spread or variability based on quantiles are the *range*, which is the distance between the minimum and maximum, and the *inter-quartil range*, the distance between the first and third quartile.

Box plots are graphical descriptions of samples based on quantiles, and sometimes other statistics (there are a number of variations).

4.5 Other

The *coefficient of variation* (CV) is σ/μ for a population, and $s/\overline{Y}$ for a sample. It is commonly given as a percentage.

As already seen, in a Poisson point process, that is, points scattered independently at random and homogeneously (constant rate) over time or space, the counts in a random plot or interval has a Poisson distribution. Both the mean and variance equal the rate parameter λ. Therefore, the *coefficient of dispersion* $s^2/\overline{Y}$ is used to assess deviations from "randomness".

5 Confidence intervals

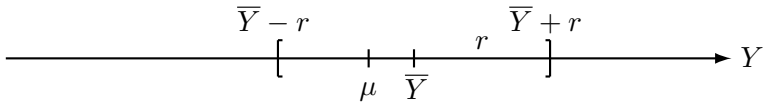
Confidence intervals are calculated so that there is a given probability of the interval containing a parameter. The ends of the interval are statistics (functions of the sample), and therefore are RVs. In Classical Statistics the parameter is an unknown but fixed number.

5.1 Interval estimation of μ, large samples

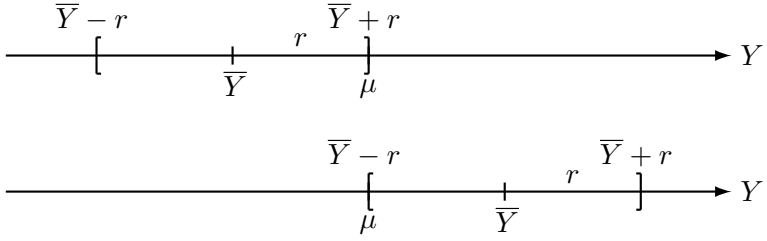
Whatever the distribution of Y, for "large" n the sample mean $\overline{Y}$ is approximately normal (Central Limit Theorem, Lab 5). Moreover, we know the mean and standard deviation (standard error), so that approximately

$$\overline{Y} \sim N(\mu, \mathrm{SE}) \ ,$$

with $\mathrm{SE} = \sigma_{\overline{Y}} = \sigma/\sqrt{n}$. Assume that s is a good enough estimate of σ, as it should be in large enough samples, so that we use the estimated SE.



We want to find a (random) interval $[\overline{Y} - r, \overline{Y} + r]$, such that the interval will contain μ with probability 0.95. We need the "radius" r.



It is seen from this picture that the interval contains μ if $\overline{Y}$ lies between $\mu - r$ and $\mu + r$. Or, if $\overline{Y} - \mu$ lies between $-r$ and r. Clearly, $\overline{Y} - \mu \sim N(0, SE)$. Let $F(y)$ be the CDF for this distribution. Then, we want an r such that

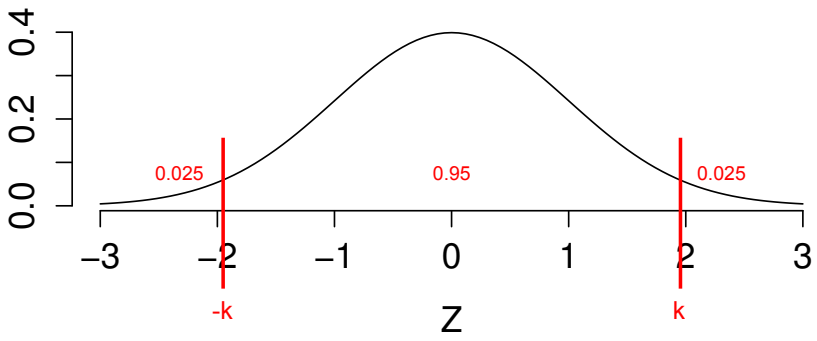
$$F(r) - F(-r) = 0.95 \; .$$

Plugging in the estimated SE, we could solve this numerically for r. In R, for instance, this can be done using `pnorm` and `uniroot`.

Traditionally, however, one goes a bit further. For more generality, and because statistical tables were only for the standard normal. We can standardize dividing by the SE:

$$\frac{\overline{Y} - \mu}{\text{SE}} = Z \sim N(0, 1) \; .$$

Making $r = k\text{SE}$, we look now for a k such that the probability of Z being between $-k$ and k is 0.95:



There are a few ways of finding k from the areas under the standard normal PDF, see the picture. For instance, the probability of values less than k must be 0.975: $F(k) = 0.975$, or $k = F^{-1}(0.975)$, where F^{-1} is the standard normal quantil function. For `qnorm(0.975, 0, 1)`, R gives `1.959964`. Remembering that $r = k\,\mathrm{SE} = 1.96\,\mathrm{SE}$, we conclude that the confidence interval is

$$[\overline{Y} - 1.96\,\mathrm{SE}, \overline{Y} + 1.96\,\mathrm{SE}]\ .$$

Or approximately, $\overline{Y} \pm 2\,\mathrm{SE}$.

5.2 Interval estimation of μ, small normal samples

Here we assume that the Y_i are normal, or at least that $\overline{Y}$ is nearly normal. Above we assumed that the estimated SE was close enough to the population SE, so that Z was standard normal. In smaller samples this may not be so good; the estimated SE is a random variable, and Z is not normal anymore.

There are three very important distributions derived from the normal in statistics:

One is the distribution of the square of a normal RV, and more generally, the distribution of a sum of squared normals. It is called the χ^2 distribution. The sample variance is (a multiple of) a sum of squared normals, and therefore $s^2 \sim \chi^2$ (I will not bother with the distribution parameters).

The second is the distribution of the ratio of a normal and the square root of a χ^2. Just like Z. It is called the *t distribution*, or *Student's t*.

The third is the distribution of a ratio or χ^2's, the *F distribution*. Used in analysis of variance, among other things.

Here we need the t. The procedure is the same as before, except that instead of k being a normal quantil, it is a t-distribution quantil, popularly called the "t-value". The t-distribution has one parameter, called the degrees of freedom, which in this case is $n-1$. If n is large, the t quantil is close to the normal 1.96, for smaller n it is slightly larger.

5.3 Bayesian interpretation

Bayesians think of μ as having a probability distribution, a prior before looking at the data, and a posterior after. A more straightforward interpretation then makes sense, μ has a given probability, e.g., 0.96, of being inside the *observed* interval. If the prior is "flat", reflecting complete prior ignorance, the numerical results are exactly the same as before. But the interpretation is different, and the interval is sometimes called a *credibility interval* to distinguish it from the classical confidence interval.