5. Bridging to Statistics

5.1 Random Samples and Statistics

This chapter is *transitional* and involves *thinking beyond* probability distributions to the use of probability for statistical inference.

All statistics begins with the *sampling method* – how the data are collected, independent or dependent (correlated) observations, fixed or random sample size, etc. Here, we consider only the most basic sampling scheme.

DEFINITION 5.1 A simple random sample (SRS) is a vector (of fixed length) of random elements that are *mutually independent* and all have the *same distribution*.

They are also described as independent and identically distributed (iid), and we will write, for example, $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$. Each random element may be a random variable or a random vector or even a random function.

In the case of sampling from a finite population, this is the same as simple random sampling *with replacement*, as that ensures independence.

More generally we are sampling from distributions, the equivalent of infinite populations. Although infinite populations may seem unrealistic, there is no question that this viewpoint gives us useful mathematical models.

Example 5.1 Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ exponential(β). Let $f(x)$ be the common marginal pdf. The joint density of the sample is

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) &= f(x_1)f(x_2)\cdots f(x_n) \\ &= \beta^{-n}\mathrm{e}^{-(x_1+\cdots+x_n)/\beta}, \quad \text{each } x_i > 0. \end{aligned}$$

Example 5.2 If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim}$ Poisson(λ) and $f(y)$ is the Poisson(λ) pmf, then their joint pmf is

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = f(y_1)f(y_2)\cdots f(y_n) = \lambda^{y_1+\cdots+y_n}\mathrm{e}^{-n\lambda}\prod_{i=1}^{n}\frac{1_{\{0,1,\ldots\}}(y_i)}{y_i!}.$$

The form of the joint pdf or pmf is important for developing optimal statistical methods, especially for *exponential families* as in the last two examples. (Recall from Chapter 3.)

Example 5.1 (cont.) The exponential distribution is asymmetric and has a moderately long right tail which means the maximum of a random sample could be fairly large.

How large?

Solution Let $M = \max(X_1, \ldots, X_n)$. Then the cdf for M is, for $t > 0$,

$$\mathsf{P}(M \leq t) = \mathsf{P}(X_1 \leq t, \ldots, X_n \leq t) = \prod_{i=1}^{n} \mathsf{P}(X_i \leq t) = (1 - \mathrm{e}^{-t/\beta})^n.$$

Thus, taking a derivative, its pdf is

$$f_M(t) = \tfrac{n}{\beta}\mathrm{e}^{-t/\beta}(1 - \mathrm{e}^{-t/\beta})^{n-1}, \quad t > 0.$$

Probabilities and quantiles are easily obtained from the cdf for M. Although messy, one can also derive the mean and variance of M.

The point here is that the *distribution of a function* of the data (M) is just as relevant to a statistical question as is the joint distribution of the data.

**** In fact, it is often not the whole sample that interests us, but rather some specific information it provides.*

Example 5.2 (cont.) From Thm. 4.10 we know that the sum of independent Poisson rvs again has Poisson distribution. In particular, the sum of n independent Poisson(λ) rvs has Poisson($n\lambda$) distribution.

Hence we can immediately make statements like $\mathsf{E}(Y_1 + \cdots + Y_n) = n\lambda$ and $\mathsf{var}(Y_1 + \cdots + Y_n) = n\lambda$.

On the other hand, merely having independence can still lead to important and useful results without ever expressing the joint pdf or pmf.

For example, it is always the case that if $X_1, \ldots, X_n$ have the same mean μ then their total $T = X_1 + \cdots + X_n$ has mean $n\mu$. And if they are independent (hence uncorrelated) with the same variance σ^2 then $\mathsf{var}(T) = n\sigma^2$, by the rules of the previous chapter.

DEFINITION 5.2 Let $X_1, \ldots, X_n$ be a random sample. A statistic is any random variable or random vector $T = T(X_1, \ldots, X_n)$ where $T(x_1, \ldots, x_n)$ is a function that *does not depend on unknown parameter values*.

In other words, a statistic is a quantity that can be computed solely from the data.

An estimator is a statistic that is used to *estimate the value* of a parameter or of a function of the parameter(s).

Note: "to estimate" means to provide a *plausible value* for the parameter given the information in the data; it does not mean to approximate or to bound.

Obvious examples of statistics are the *sum* $X_1 + \cdots + X_n$ and the *maximum* $\max(X_1, \ldots, X_n)$. The *sample mean* $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ (average) is frequently used to estimate the distribution mean.

We will investigate a number of other statistics and estimators as well.

Note: this chapter is focused on illustrating the *role of probability* in statistical inference. Further and more general theory is for another course.

**** We are now at a critical spot in the course.*

To appreciate the rest of this chapter we need to consider the perspective of a statistician looking for useful and acceptable ways to understand data.

Indeed, we must simultaneously take two opposing perspectives: *looking forward* to the (random) sample, its statistics and the methods of inference while also, having data in hand, *looking back* to the model and to what the data say about the model.

CONCEPT 5.3 Statistical inference (recall Con. 1.3) involves interpreting data in the context of a probability model where the parameters have *unknown values*.

Specifically, statistical estimation is a method for using the data (sample) values in order to guess the parameter value. Estimation necessarily involves the use of statistics (as in Def. 5.2), the quality of which depends on their own probability distributions.

Henceforth, therefore, we proceed as if we know the model and we will refer to the parameter(s), but we *assume we do not know the actual value(s)* of the parameter(s).

Since one purpose of statistics is to estimate parameters of the distribution, and since the *moments* of the distribution are among the most important parameters, it is natural that we would be interested in the statistics that do this estimation.

DEFINITION 5.4 Let $X_1, \ldots, X_n$ be a random sample.

i. The sample mean is $\overline{X} = \widehat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} X_i$.

ii. The sample k-th moment is $\widehat{\mu}'_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$.

iii. The sample k-th central moment is $\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$.

iv. The sample variance is $\widehat{\sigma}^2_X = \widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

v. The sample standard deviation $\widehat{\sigma}_X = \sqrt{\widehat{\sigma}^2_X}$.

Each of these estimates the corresponding (true) value of the distribution parameter, assuming it is finite.

**** Be careful to notationally distinguish an estimator from the parameter, for example by use of the "hat".*

- For example, $\overline{X}$ estimates the distribution mean $\mu_X = \mathsf{E}(X)$, $\widehat{\mu}_2'$ estimates $\mu_2' = \mathsf{E}(X^2)$ and $\widehat{\sigma}_X^2$ estimates $\sigma_X^2 = \mathsf{var}(X)$.
- The *conventional* definition for sample variance is slightly different but it does not satisfy all the properties of a variance, so for now we are using the *natural* definition instead. More about this later.
- Each of the estimators above are valid (and natural) as long as the corresponding parameters are finite. But they are not necessarily the *optimal* estimators, which depend on the particular distribution model.

Example 5.2 (cont.) $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. The sample mean, $\overline{Y}$ is a natural estimator of the true mean λ. Indeed,

$$\mathsf{E}(\overline{Y}) = \frac{1}{n}\mathsf{E}(Y_1 + \cdots + Y_n) = \lambda,$$

making $\overline{Y}$ unbiased. (That is, the expectation of the statistic is the value it is estimating, for all values of the parameter).

Of course, for Poisson data, the sample variance $\widehat{\sigma}^2_Y$ is *also* a reasonable estimator for λ, since λ is the true variance as well as being the mean.

The question then is whether (or under what circumstances) ought we use $\overline{Y}$ to estimate λ, or $\widehat{\sigma}^2_Y$, or perhaps something else is even better.

Example 5.1 (cont.) $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ exponential(β). The sample mean, $\overline{X}$ estimates the true mean β and, again,

$$\mathsf{E}(\overline{X}) = \frac{1}{n}\mathsf{E}(X_1 + \cdots + X_n) = \beta.$$

But the sample standard deviation $\widehat{\sigma}_Y$ is also a reasonable estimator for β, since β is the true standard deviation.

Again, we have the question which is best, if either.

**** These two examples show there can easily be more than one reasonable way to estimate a parameter.*

The point is this: unless we understand the estimators (statistics) from a probability point of view by studying the properties of their distributions, we have no hope of knowing how *best* to do statistical inference.

DEFINITION 5.5 Suppose $X_1, \ldots, X_n$ is a random sample and $T = T(X_1, \ldots, X_n)$ is a function of the sample. (It could be a statistic or it could also depend on parameter values.)

The probability distribution for T is called the sampling distribution for T, meaning it is the distribution of T among the sample space of random samples.

This is a distribution just as for any other random variable. The term "sampling" merely reminds us that the random variable is the result of a random sample.

Example 5.1 (cont.) For exponential(β) data, the sum has gamma(n, β) distribution.

It follows, by a transformation, that $\overline{X}$ has gamma($n, \beta/n$) distribution.

Moreover, $\overline{X}/\beta \sim$ gamma$(n, 1/n)$. Thus the distribution of $\overline{X}/\beta$ does not depend on the parameter, although the rv itself does. (This fact can be useful.)

Example 5.3 Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{normal}(\mu, \sigma^2)$. We know that sums, and indeed linear combinations, of independent normal rvs have normal distribution.

Using the expectation and variance formulas for sums, we can thus show that $\overline{X} \sim \text{normal}(\mu, \sigma^2/n)$ and $\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim \text{normal}(0, 1)$ (exercise).

- Although estimators and statistics are computed without knowing the parameter values, their distributions usually do depend on the parameter values.
- Actually, this is necessary for them to be relevant to statistical inference since otherwise we would have no way to relate them to the quantities they are supposed to estimate.
- On the other hand, there often are random variables that depend on both the data and the parameters in such a way that their distributions do not depend on the parameters.
- Such rvs, known as pivots, are very useful for developing statistical methods.

5.2 Sums, Means and Moments

Essentially, mathematical statistics is *the study of sampling distributions* for important and useful statistical estimators and such.

Clearly, the first to be studied are the sampling distributions for the total of a random sample, and for $\overline{X}$ and other sample moments.

We have seen how the convolution formula enables us to determine the distribution of the sum of independent rvs. Sometimes this is relatively easy (as in the case of Poisson rvs), but usually it gets to be very hard when the number of independent rvs is arbitrary or large.

Moment generating functions can also be useful in this regard. Computation is easy – assuming you have the mgf in the first place. Additionally, they are also very useful for certain proofs, as we will see.

THEOREM 5.6 Suppose $X_1, X_2, \ldots, X_n$ are *independent* random variables with mgfs $M_1, M_2, \ldots, M_n$, respectively.

Then $X_1 + \cdots + X_n$ has mgf $M(t) = \prod_{i=1}^{n} M_i(t)$.

In particular, if $X_1, X_2, \ldots, X_n$ are iid then $X_1 + \cdots + X_n$ has mgf $M(t) = (M_1(t))^n$ and $\overline{X}$ has mgf $M_{\overline{X}}(t) = (M_1(t/n))^n$.

PROOF By independence, the mgf of $X_1 + \cdots + X_n$ is

$$\mathsf{E}(\mathrm{e}^{t(X_1+\cdots+X_n)}) = \mathsf{E}(\mathrm{e}^{tX_1} \cdots \mathrm{e}^{tX_n}) = M_1(t) \cdots M_n(t).$$

If the rvs are *iid* then they all have the same mgf, $M_1(t)$, and thus the mgf of $X_1 + \cdots + X_n$ is $(M_1(t))^n$.

By Thm. 2.28, the mgf of $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ is then $(M_1(t/n))^n$. □

**** This theorem, in combination with Thm. 2.29.ii, can sometimes be used to identify the distribution of a sum of independent rvs without ever computing a convolution.*

Example 5.4 Suppose $X \sim$ binomial(m, p) and $Y \sim$ binomial(n, p), independent.

Then $M_X(t) = (pe^t + 1 - p)^m$ and $M_Y(t) = (pe^t + 1 - p)^n$. (See Ex. 2.5.)

Therefore,

$$M_{X+Y}(t) = M_X(t)M_Y(t) = (pe^t + 1 - p)^{m+n},$$

which says $X + Y \sim$ binomial$(m + n, p)$. Note: same p for both.

Example 5.2 (cont.) If $Y_1, \ldots, Y_n \sim$ Poisson(λ) then the mgf of Y_1 is $M_1(t) = e^{\lambda(e^t-1)}$.

Thus, the mgf of $Y_1 + \cdots + Y_n$ is

$$M(t) = (e^{\lambda(e^t-1)})^n = e^{n\lambda(e^t-1)},$$

which is the mgf for the Poisson$(n\lambda)$ distribution.

The mgf for $\overline{Y}$ can also be found: $M_{\overline{Y}}(t) = e^{n\lambda(e^{t/n}-1)}$. Of course, the support for $\overline{Y}$ is $0, 1/n, 2/n, \ldots$, so its distribution is not one of the standard discrete distributions.

Example 5.3 (cont.) Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{normal}(\mu, \sigma^2)$.

The mgf of X_1 is $M_1(t) = \mathrm{e}^{\mu t + \sigma^2 t^2/2}$ and so the mgf of $\overline{X}$ is

$$M_{\overline{X}}(t) = (M_1(t/n))^n = \left(\mathrm{e}^{\mu t/n + \sigma^2 (t/n)^2/2}\right)^n = \mathrm{e}^{\mu t + (\sigma^2/n) t^2/2}.$$

This shows that $\overline{X} \sim \text{normal}(\mu, \sigma^2/n)$.

Now, to put this into a statistical context: A certain crab species is known to have average body mass $\mu = 57$g and a population standard deviation $\sigma = 11$g. The distribution of body mass can be modeled with a normal distribution.

A biologist finds what appears to be a new subspecies at a new location. From a sample of 18 crabs he obtains an average $\overline{X} = 67$g.

Is this reason enough to believe he has actually found a new subspecies?

Solution Here, $n = 18$. We know $\overline{X} \sim \text{normal}(\mu, \sigma^2/n)$. This in turn says that $\sqrt{n}(\overline{X} - \mu)/\sigma$ has normal(0,1) distribution. Thus, if the new crabs are in fact the original (sub)species,

$$\begin{aligned} \mathsf{P}(\overline{X} \geq 67) &= \mathsf{P}(\sqrt{n}(\overline{X} - \mu)/\sigma \geq \sqrt{18}(67 - 57)/11) \\ &= 1 - \Phi(3.857) \doteq .0000574. \end{aligned}$$

This is extremely small and says the finding (67g) is very surprising if the new crabs are the same as the others. One may therefore conclude that they are a different subspecies and then estimate their mean body mass to be 67g.

The point here is that the (sampling) distribution of $\overline{X}$ is used to answer the statistical question.

Example 5.5 (Chi-Square RVs) Suppose $Z_1, \ldots, Z_m$ are iid standard normal random variables. We saw earlier, in Thm. 3.11, that $Z_i^2 \sim$ chi-square(1) which is the same as gamma$(1/2, 2)$.

Since $Z_1^2, \ldots, Z_m^2$ are also independent, their sum $S = Z_1^2 + \cdots + Z_m^2$ is a chi-square(m) (gamma$(m/2, 2)$) random variable. (Exercise – use gamma mgfs. Also see Ex. 4.12 for another proof.)

We also have, for $S \sim$ chi-square(m), that E$(S) = m$ and var$(S) = 2m$.

This is a *very important example* as many statistics useful for hypothesis testing are either the sum of squared normal rvs or can be approximated by such a sum.

Example 5.6 The mgf for the Laplace(μ, β) distribution is

$$M_V(t) = \mathrm{e}^{\mu t}(1 - \beta^2 t^2)^{-1}.$$

Suppose $V_1, \ldots, V_n$ is a random sample from this distribution. The mgf of the sample mean $\overline{V}$ is thus

$$\begin{aligned} M_{\overline{V}}(t) = (M_V(t/n))^n &= \left(\mathrm{e}^{\mu t/n}(1 - \beta^2(t/n)^2)^{-1}\right)^n \\ &= \mathrm{e}^{\mu t}(1 - \beta^2 t^2/n^2)^{-n}, \end{aligned}$$

which is not especially informative (at this point at least).

However, we recall that the Laplace distributions are a location-scale family and let $W_i = \frac{V_i - \mu}{\beta}$, $i = 1, \ldots, n$. It is easy to see then that $\overline{W} = \frac{\overline{V} - \mu}{\beta}$. Since the W_i's have Laplace$(0, 1)$ distribution, we may conclude $\overline{W}$ has mgf

$$M_{\overline{W}}(t) = (1 - t^2/n^2)^{-n}.$$

This still does not tell us how to compute probabilities but it is a little cleaner and, more importantly, it shows us how the statistic $\overline{V}$ and the parameters are related. (The distribution of $\frac{\overline{V} - \mu}{\beta}$ does not depend on μ or β.)

Ideally, we can state exactly what the sampling distribution is (either of the statistic itself or of a simple 1-1 function of the statistic).

This is not generally possible, however, so we at least hope to say something about the properties of the estimators.

Sums and averages, in particular, have properties that we can easily describe.

THEOREM 5.7 Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, for some (marginal) distribution F. Then, subject to existence, the following hold.

i. $\mathsf{E}(\overline{X}) = \mu = \mathsf{E}(X_1)$ and $\mathsf{var}(\overline{X}) = \frac{1}{n}\mathsf{var}(X_1)$. So $\overline{X}$ is *unbiased* for μ.

ii. $\mathsf{E}(\widehat{\mu}'_k) = \mu'_k = \mathsf{E}(X_1^k)$ and $\mathsf{var}(\widehat{\mu}'_k) = \frac{1}{n}\left(\mu'_{2k} - (\mu'_k)^2\right)$. $\widehat{\mu}'_k$ is *unbiased* for μ'_k.

Note that very little is assumed about the distribution F.

PROOF Since they are identically distributed, each X_i has mean $\mathsf{E}(X_1)$ and variance $\mathsf{var}(X_1)$. Thus,

$$\mathsf{E}(\overline{X}) = \mathsf{E}\left(\frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n\right) = \mathsf{E}(X_1).$$

By the variance formula for linear combinations of independent rvs,

$$\begin{aligned}\mathsf{var}(\overline{X}) &= \mathsf{var}\Big(\frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n\Big) \\ &= \frac{1}{n^2}\mathsf{var}(X_1) + \cdots + \frac{1}{n^2}\mathsf{var}(X_n) = \frac{1}{n}\mathsf{var}(X_1).\end{aligned}$$

Observe that i. is a special case of ii. with $k = 1$.

But at the same time, ii. is an example of i. Specifically, if $Y_i = X_i^k$ then $Y_1, \ldots, Y_n$ is also an iid random sample. Therefore, we can apply i. to the sample $Y_1, \ldots, Y_n$ using the facts

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} X_i^k = \widehat{\mu}'_k,$$

$$\mathsf{E}(Y_1) = \mathsf{E}(X_1^k) = \mu'_k \quad \text{and} \quad \mathsf{var}(Y_1) = \mathsf{E}((X_1^k)^2) - (\mathsf{E}(X_1^k))^2 = \mu'_{2k} - (\mu'_k)^2.$$

This proves ii. □

**** In fact the same argument applies to the average of any function of the data, such as* $\frac{1}{n}\Big(g(X_1) + \cdots + g(X_n)\Big)$*, making this an especially useful result.*

- Averages, such as the sample moments, are generally unbiased estimators. However, not all estimators have that property, including transformations of averages.
- On the other hand, useful estimators should have a relatively small bias, particularly when the number of data (n) is *large*.
- Similarly, we generally want the variance to get smaller as the number of data increases. This is something we will investigate in the next section.
- The standard error of an estimator is the standard deviation of its sampling distribution. This term is used *just for estimators*.
- While the expectation and variance of a statistic are highly useful, they can be misleading if the sampling distribution is *asymmetric or heavy-tailed*. Fortunately, as we will see, things improve for larger sample sizes.

Central moments can likewise be studied, especially the sample variance.

THEOREM 5.8 Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, for some (marginal) distribution F. Then, subject to existence, the following hold.

i. $\mathsf{E}(\widehat{\sigma}_X^2) = \left(1 - \frac{1}{n}\right)\sigma_X^2$.

ii. $\mathsf{var}(\widehat{\sigma}_X^2) = \frac{1}{n}(1 - \frac{1}{n})^2(\mu_4 - (1 - \frac{2}{n-1})\sigma_X^4)$.

iii. $\mathsf{cov}(\widehat{\sigma}_X^2, \overline{X}) = \frac{1}{n}(1 - \frac{1}{n})\mu_3$.

PROOF i. Since $\widehat{\sigma}_X^2$ is a (true) variance, we have $\widehat{\sigma}_X^2 = \widehat{\mu}_2' - (\overline{X})^2$ (check). Therefore,

$$\begin{aligned}\mathsf{E}(\widehat{\sigma}_X^2) &= \mathsf{E}(\widehat{\mu}_2') - \mathsf{E}((\overline{X})^2) = \mu_2' - (\mathsf{var}(\overline{X}) + \mu_X^2) \\ &= \mathsf{var}(X_1) - \frac{1}{n}\mathsf{var}(X_1) = \left(1 - \frac{1}{n}\right)\sigma_X^2.\end{aligned}$$

ii. and iii. (exercises. It helps to work with the centered random variables $Y_i = X_i - \mu_X$ as they have mean 0.) □

- Because Thm. 5.8.i. indicates that $\widehat{\sigma}_X^2$ is not quite unbiased, it is *conventional* to modify the sample variance to $S^2 = \frac{n}{n-1}\widehat{\sigma}_X^2$, which is unbiased. That is, $\mathsf{E}(S^2) = \sigma_X^2$.
- The estimate for σ_X, however, *remains biased*: $0 < \mathsf{var}(S) = \mathsf{E}(S^2) - (\mathsf{E}(S))^2 = \sigma_X^2 - (\mathsf{E}(S))^2$; hence $\mathsf{E}(S) < \sigma_X$.
- The other properties in Thm. 5.8 can be adjusted similarly to get $\mathsf{var}(S^2)$ and $\mathsf{cov}(\overline{X}, S^2)$.
- Note that $\overline{X}$ and $\widehat{\sigma}^2$ are uncorrelated only if the skewness coefficient μ_3/σ^3 is zero. It turns out that this is the reason the t-test (see Sec. 5.4) does not work as well for asymmetric data as it does for symmetric data.

We can also find expectation and variance of an estimator for covariance.

THEOREM 5.9 Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a simple random sample of bivariate data. Define the covariance estimator

$$\widehat{\text{cov}}_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}).$$

i. $\text{E}(\widehat{\text{cov}}_{X,Y}) = \left(1 - \frac{1}{n}\right)\text{cov}(X, Y)$.

ii. $\text{var}(\widehat{\text{cov}}_{X,Y}) = \frac{1}{n}(1 - \frac{1}{n})^2(\text{E}((X - \mu_X)^2(Y - \mu_Y)^2) - (1 - \frac{2}{n-1})\sigma_X^2\sigma_Y^2)$.

Note, however, that the more conventional estimator for covariance uses divisor $n - 1$ in place of n, similar to S^2, and this is then *unbiased* for $\text{cov}(X, Y)$.

PROOF i. (only) First we observe that, with some manipulation,

$$\widehat{\text{cov}}_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) - (\overline{X} - \mu_X)(\overline{Y} - \mu_Y).$$

The first term is an average, with mean $\text{E}((X_1 - \mu_X)(Y_1 - \mu_Y)) = \text{cov}(X, Y)$. The second term has expectation $\text{cov}(\overline{X}, \overline{Y}) = \frac{1}{n}\text{cov}(X, Y)$, using an extension to Thm. 4.34.i. □

In general, of course, the sampling distribution of a statistic is very complicated, even if that statistic is just the sample mean. Among other things, it depends heavily on the distribution that the data come from.

Example 5.7 10 observations are sampled (without replacement) from a very small population, and the sample mean and standard deviation are computed.

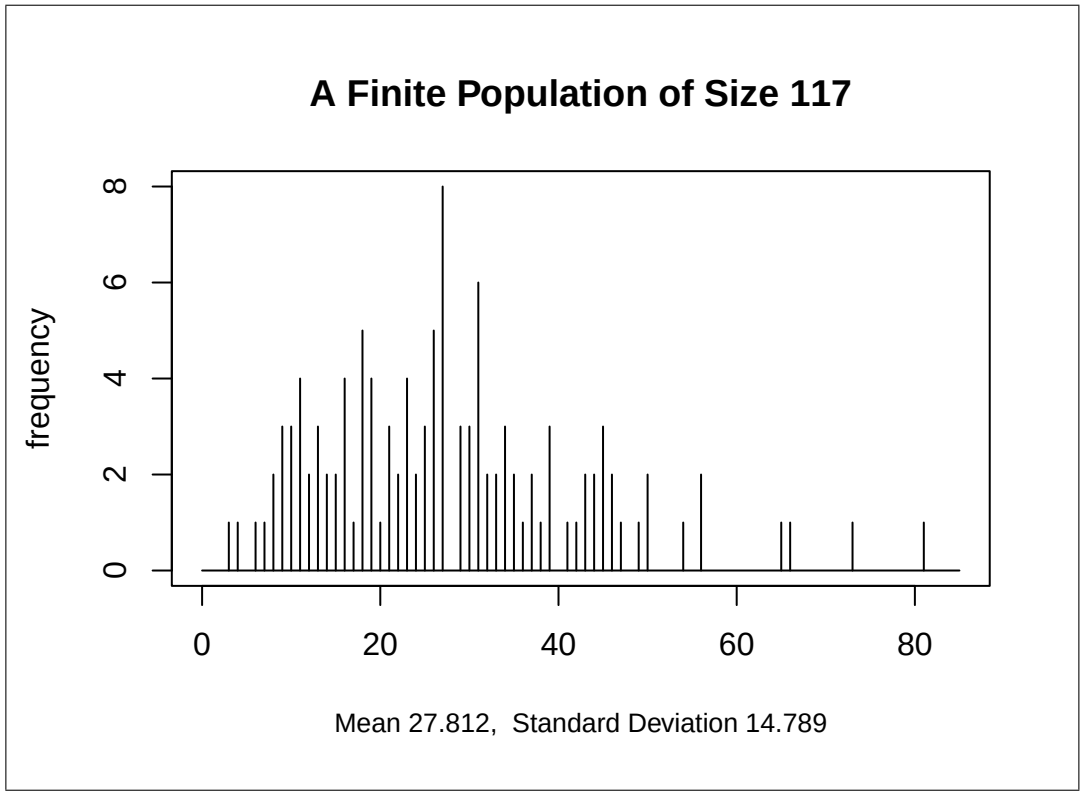

Figure 5.1 Line graph for values of a small population.

What do the sampling distributions for those statistics look like?

Solution Since we have the population at hand, we can simulate many random samples, calculate the statistics from each sample, and then observe the distributions of those statistics.

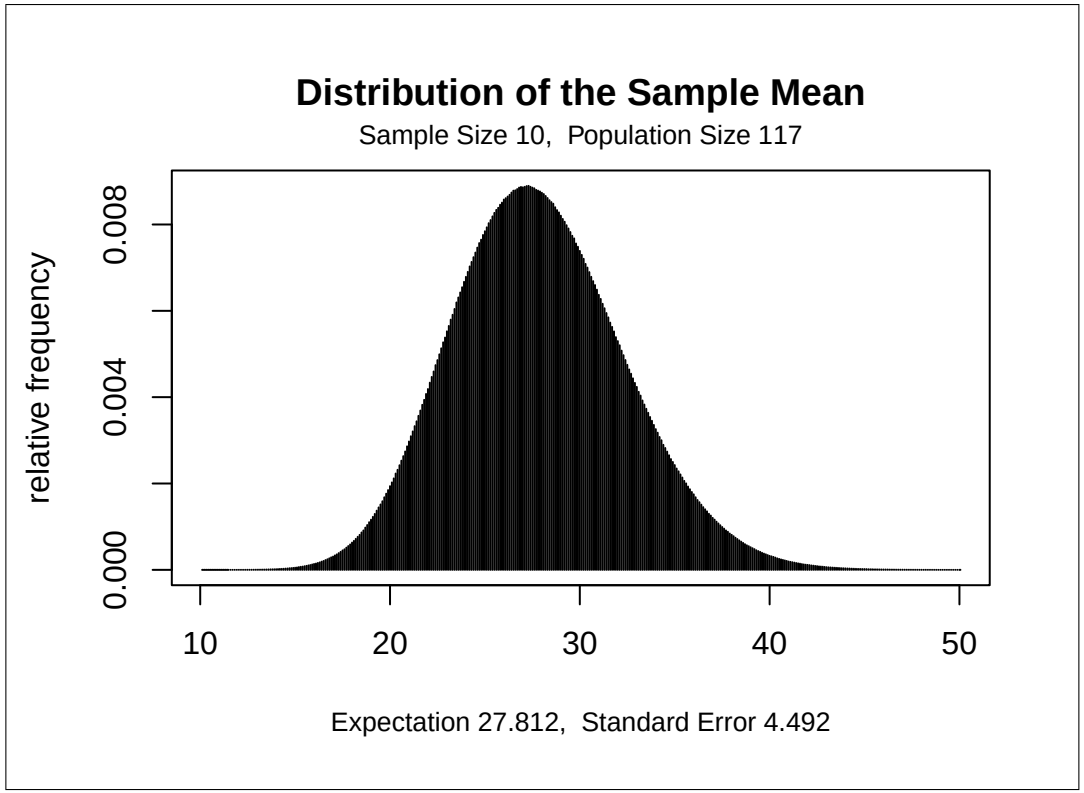

Figure 5.2 Histogram for 100 million simulated values of the sample average, when $n = 10$ (sampled without replacement).

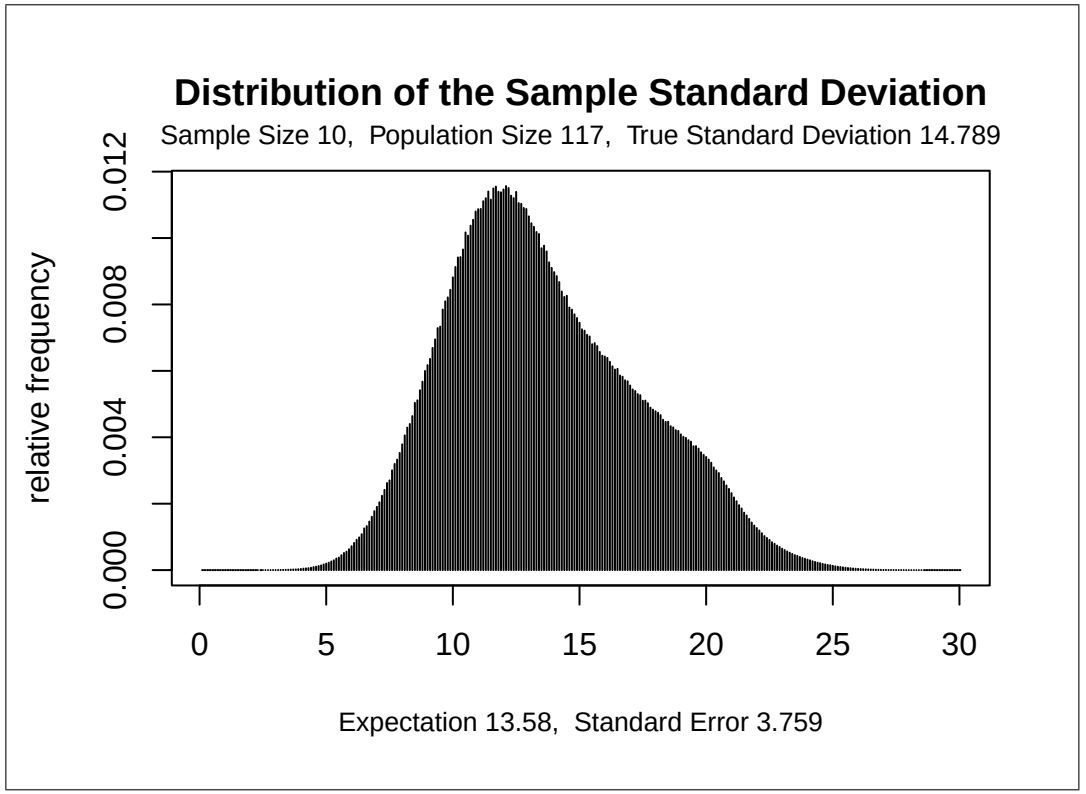

Figure 5.3 Histogram for 100 million simulated values of the sample standard deviation, when $n = 10$ (sampled without replacement).

The number of *possible* random samples of size 10 is of course

$$\binom{117}{10} \approx 8.92 \times 10^{13}.$$

5.3 Statistical Limit Theorems

There are two main reasons for considering "large sample" theory. The first is to ensure that our statistics *get better with more data*. The second is to *approximate sampling distributions* that otherwise are impossible to compute.

We start with a discussion of the first objective. This requires defining what it means for random variables to converge.

DEFINITION 5.10 Suppose $T_1, T_2, T_3, \ldots$ is a sequence of (typically *dependent*) rvs, defined on the same sample space and T is another rv defined on that sample space.

i. T_n converges in probability to T, as $n \to \infty$, if $\mathsf{P}(|T_n - T| \leq \delta) \to 1$ for any $\delta > 0$. We write $T_n \to T$ i.p.

ii. T_n converges in mean square to T, as $n \to \infty$, if $\mathsf{E}(|T_n - T|^2) \to 0$. We write $T_n \to T$ in m.s. Also, T_n converges in mean to T if $\mathsf{E}(|T_n - T|) \to 0$.

iii. T_n converges with probability 1 to T, as $n \to \infty$, if $\mathsf{P}(\lim_{n\to\infty} T_n = T) = 1$. We write $T_n \to T$ w.p. 1.

Essentially the same definitions apply for sequences of random vectors, also.

**** Although each of these convergence types indicate that the sequence of rvs T_n is getting closer to T, they are in fact different definitions. Each has important uses in probability theory and statistics.*

THEOREM 5.11 Suppose as in Def. 5.10.

i. $T_n \to T$ i.p. is implied by either $T_n \to T$ in mean or in m.s. or w.p. 1.

ii. $T_n \to T$ in m.s. implies $T_n \to T$ in mean.

iii. Neither of $T_n \to T$ in m.s. or $T_n \to T$ w.p. 1 implies the other.

PROOF The proof of most of this theorem is beyond the scope of this course, but we can show part of i. Suppose $T_n \to T$ in m.s. By applying Markov's inequality (Thm. 2.23),

$$\mathsf{P}(|T_n - T| > \delta) \leq \frac{\mathsf{E}(|T_n - T|)}{\delta} \to 0.$$

Thus

$$\mathsf{P}(|T_n - T| \leq \delta) = 1 - \mathsf{P}(|T_n - T| > \delta) \to 1$$

so that $T_n \to T$ i.p.

Part ii. follows from the fact $(\mathsf{E}(|T_n - T|))^2 \leq \mathsf{E}(|T_n - T|^2)$. □

The most important instance of convergence is for a mean. For this, we expand our notion of sample space to allow for sequences of random variables.

THEOREM 5.12 Assume $X_1, X_2, \ldots$ is a sequence of *iid* rvs. Define $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, the mean of the sample $(X_1, \ldots, X_n)$, for each $n \geq 1$.

i. (Strong Law of Large Numbers) If $\mathsf{E}(|X_1|) < \infty$ then $\overline{X}_n \to \mathsf{E}(X_1)$ with probability 1 and $\mathsf{E}(|\overline{X}_n - \mathsf{E}(X_1)|) \to 0$ (convergence in mean).

ii. If $\mathsf{E}(|X_1|^2) < \infty$ then $\overline{X}_n \to \mathsf{E}(X_1)$ in mean square.

PROOF ii. (only) We have

$$\mathsf{E}(|\overline{X}_n - \mathsf{E}(X_1)|^2) = \mathsf{var}(\overline{X}_n) = \frac{\mathsf{var}(X_1)}{n} \to 0,$$

as $n \to \infty$. □

**** Recall that $\mathsf{E}(|X_1|^2) < \infty$ implies $\mathsf{E}(|X_1|) < \infty$ by Thm. 2.24.i.*

For statisticians, the primary issue is whether a given method of estimation is converging as the sample size increases. By Thm. 5.12, we see that this is so for the sample mean. Statisticians therefore say that the sample mean is a consistent estimator for $\mathsf{E}(X_1)$.

Example 5.2 (cont.) Let $Y_1, Y_2, \ldots$ be an iid sequence of Poisson(λ) rvs. Since $\mathsf{E}(Y_1^2) < \infty$, we find that $\overline{Y}_n$ converges both in mean square and with probability 1 to $\lambda = \mathsf{E}(Y_1)$, as $n \to \infty$. $\overline{Y}_n$ is thus consistent for λ.

Example 5.3 (cont.) Let $X_1, X_2, \ldots$ be an iid sequence of normal(μ, σ^2) rvs. Since $\mathsf{E}(X_1^2) < \infty$, we again find that $\overline{X}_n$ converges in mean square and with probability 1 and is consistent for the true mean μ.

Now let $W_n = X_n^2$, for $n \geq 1$. The sample second moment is $\widehat{\mu}_2' = \frac{1}{n}\sum_{i=1}^n X_i^2 = \overline{W}$ (which depends on $X_1, \ldots, X_n$). Since $\mathsf{E}(W_1^2) = \mathsf{E}(X_1^4) < \infty$, $\widehat{\mu}_2'$ converges both in mean square and with probability 1 to $\mathsf{E}(X_1^2) = \mu^2 + \sigma^2$, as $n \to \infty$, and is thus consistent for $\mu^2 + \sigma^2$.

Since $\widehat{\sigma}^2 = \widehat{\mu}_2' - (\overline{X})^2$, we might expect that it is consistent for $\mu_2' - \mu^2 = \sigma^2$. Indeed this is so. First, we already know that $\mathsf{E}(\widehat{\sigma}^2) = (1 - \frac{1}{n})\sigma^2$ and $\mathsf{var}(\widehat{\sigma}^2) \to 0$ by Thm. 5.8. Thus, with a bit of algebra,

$$\mathsf{E}((\widehat{\sigma}^2 - \sigma^2)^2) = \mathsf{var}(\widehat{\sigma}^2 - \sigma^2) + (\mathsf{E}(\widehat{\sigma}^2 - \sigma^2))^2 = \mathsf{var}(\widehat{\sigma}^2) + \frac{\sigma^4}{n^2} \to 0.$$

This says $\widehat{\sigma}^2 \to \sigma^2$ in mean square. Second, the strong (w.p. 1) convergence also holds, by applying part i. of the following result with $T_n = (\widehat{\mu}_2', \overline{X})$ and $g(u, v) = u - v^2$.

In fact, applying a continuous function to the members of a converging random sequence does exactly what you would expect.

THEOREM 5.13 Suppose $T_1, T_2, \ldots$ is a sequence of random variables (or vectors) that converges to T with probability 1.

i. If $g(t)$ is continuous then $g(T_n) \to g(T)$ w.p. 1.

ii. If, in addition, $\max_n \mathsf{E}(|g(T_n)|^{2+\delta}) < \infty$ for some $\delta > 0$ then $g(T_n) \to g(T)$ in m.s. also.

- We have seen that $\overline{X}$ is consistent whatever the distribution, if its mean and variance exist.
- In fact it is true of all the sample moments and sample central moments as long as the required expectations exist. That is,

$$\widehat{\mu}'_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \to \mathsf{E}(X_1^k) = \mu_k,$$

in mean and with probability 1 (and in mean square if $\mu_{2k} < \infty$). Moreover,

$$\widehat{\sigma}^2 = \widehat{\mu}'_2 - (\widehat{\mu}'_1)^2 \to \mu'_2 - (\mu'_1)^2 = \sigma^2.$$

- Applications of the theorem therefore include consistency of the sample standard deviation,

$$\widehat{\sigma} = \sqrt{\widehat{\sigma}^2} \to \sigma,$$

the coefficient of variation, and the skewness coefficient, all of which can be expressed as continuous functions of the basic sample moments.

- Other examples are the ratio of two sample means and the ratio of two sample variances from data with possibly different distributions.

The second objective of large sample theory is *to approximate* the distributions of rvs in a sequence. This is a very deep topic, but we can focus on the situations that are the most useful for statistics.

Recall the notion of convergence of distributions given in Thm. 2.32: *the cdfs converge* except possibly at the jump points of the limiting distribution.

Furthermore, recall that Thm. 2.32 also indicated we can prove convergence in distribution *by showing the mgfs converge*. We do just that here in giving the most universally important theorem of probability and statistics.

THEOREM 5.14 (Central Limit Theorem or CLT) Assume $X_1, X_2, \ldots$ is a sequence of iid rvs and $\mathsf{E}(X_1^2) < \infty$. Let $\mu = \mathsf{E}(X_1)$ and $\sigma^2 = \mathsf{var}(X_1)$. Define $\overline{X}_n$ as in Thm. 5.12.

The sequence of rvs $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ converges *in distribution* to the standard normal distribution $\Phi(z)$. That is, for any real z,

$$\mathsf{P}\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq z\right) \to \Phi(z), \quad \text{as } n \to \infty.$$

This is also denoted $\sqrt{n}(\overline{X}_n - \mu)/\sigma \xrightarrow{\mathrm{D}} \mathsf{normal}(0, 1)$.

PROOF The theorem holds generally, but we restrict our proof to the case that the mgfs exist.

Let $M_X(t)$ be the common mgf for the X_i's. Put $Z_i = \frac{X_i-\mu}{\sigma}$ and $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i = \frac{\overline{X}_n-\mu}{\sigma}$. We know that the mgf for the Z_i's is $M_Z(t) = \mathrm{e}^{-\mu t/\sigma} M_X(t/\sigma)$ and $M_{\overline{Z}_n}(t) = (M_Z(t/n))^n$.

We also know $M_Z'(0) = \mathsf{E}(Z_1) = 0$ and $M_Z''(0) = \mathsf{E}(Z_1^2) = 1$.

We need to study the mgf for $\sqrt{n}\overline{Z}_n$:

$$M_{\sqrt{n}\overline{Z}_n}(t) = M_{\overline{Z}_n}(\sqrt{n}t) = (M_Z(t/\sqrt{n}))^n.$$

Since $t/\sqrt{n} \to 0$ as $n \to \infty$, we may use *Taylor's expansion* to get

$$\begin{aligned} M_{\sqrt{n}\overline{Z}_n}(t) &= \left(1 + M_Z'(0)(t/\sqrt{n}) + \tfrac{1}{2}M_Z''(0)(t/\sqrt{n})^2 + \cdots\right)^n \\ &\doteq \left(1 + \frac{t^2}{2n}\right)^n \to \mathrm{e}^{t^2/2}, \end{aligned}$$

where we have used Thm. 2.33 for the last step.

We recognize the limit as the *standard normal mgf*. So by Thm. 2.32 we conclude that $\sqrt{n}\overline{Z}_n$ converges in distribution to the normal(0,1) distribution, and this is the conclusion we wanted. □

**** Despite the central limit theorem, the random variables $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ do not converge in probability to any rv or constant. The CLT is only about their distributions.*

- The conclusion of the CLT is *equivalent to*

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\text{D}} \text{normal}(0, \sigma^2).$$

  Why? (Think about how σ is a scale parameter for normal distributions.)

- Note that we proved a special case of the CLT in Ex. 2.19.
- The SLLN tells us $\overline{X}_n$ is consistent, but the CLT tells us *how to predict* $\overline{X}_n$ with probability statements.
- It is common to refer to the asymptotic distribution of a statistic or estimator. This is the approximate distribution derived from a limit theorem such as the CLT. So, for example, we say $\overline{X}_n$ is *asymptotically* normal$(\mu, \frac{\sigma^2}{n})$.
- However, *do not say* $\overline{X}_n$ converges to normal$(\mu, \frac{\sigma^2}{n})$ as that *is not a limit*, and what $\overline{X}_n$ converges to is simply μ.

- Like Thm. 5.12, there are no distributional requirements for the central limit theorem other than existence of the needed moments. This means it may be applied quite generally.
- For example, it may be applied to sample moments. Consider an iid sequence $X_1, X_2, \ldots$ and let $Y_i = X_i^2$. Then $\overline{Y}_n = \widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$ is the sample second moment for the X_i data. If $\mathsf{E}(X_i^4) = \mathsf{E}(Y_i^2) < \infty$ then we can apply the CLT to $\overline{Y}_n$ using the mean and variance of Y_i. In other words,

$$\sqrt{n}(\widehat{\mu}_2' - \mu_2') \xrightarrow{\mathrm{D}} \mathsf{normal}(0, \mu_4' - (\mu_2')^2).$$

- However, depending on the distribution, there could be *better* approximations.

Example 5.6 (cont.) Suppose we have an iid sequence $X_1, X_2, \ldots$ from the Laplace(μ, β) distribution. The mean is μ, and the variance is $\sigma_X^2 = 2\beta^2$. By the CLT, we have that $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{2}\beta}$ converges in distribution to the normal(0,1) distribution.

For a statistical application: assume that a device that measures infrared radiation has a measurement error that can be modeled by a Laplace $(0, 0.1)$ distribution. If twenty measurements are taken, what is the chance the average error is more than .05 in size?

Solution Let $\overline{X}$ be the average error of the 20 measurements. Then $\mathsf{E}(\overline{X}) = 0$ and, assuming the measurements are independent, the standard deviation for $\overline{X}$ is $\sqrt{2}(0.1)/\sqrt{20} = .031623$.

Furthermore, the CLT suggests that the distribution of $\overline{X}$ is approximately normal if 20 is a "large enough" sample size. We will assume that it is.

Therefore,

$$\begin{aligned} \mathsf{P}(|\overline{X}| > .05) &= 1 - \mathsf{P}\left(\frac{-.05}{.031623} \leq \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{2}\beta} \leq \frac{.05}{.031623}\right) \\ &\doteq 1 - \Phi(1.58113) + \Phi(-1.58113) = .114. \end{aligned}$$

An *extremely important* example comes next.

Example 5.8 Let Y_n be the number of successes in a sample of n independent Bernoulli(p) trials. Thus Y_n is the total of n iid rvs with mean p and variance $p(1-p)$.

The *sample proportion* of successes, $\widehat{p}_n = Y_n/n$, is the average (sample mean) of the *trials*.

By the CLT, $\frac{\sqrt{n}(\widehat{p}_n - p)}{\sqrt{p(1-p)}}$ converges in distribution to the standard normal.

Since $\widehat{p} \to p$ w.p. 1 then $\sqrt{\widehat{p}_n(1-\widehat{p}_n)} \to \sqrt{p(1-p)}$ w.p. 1 by Thm. 5.13.

Thus, $\frac{\sqrt{n}(\widehat{p}_n - p)}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}$ also converges in distribution to the standard normal distribution by Slutsky's Theorem below.

We know, of course, that $\mathsf{E}(Y_n) = np$ and $\mathsf{var}(Y_n) = np(1-p)$. It is easy to see that

$$\frac{Y_n - np}{\sqrt{np(1-p)}} = \frac{\sqrt{n}(\widehat{p} - p)}{\sqrt{p(1-p)}}.$$

So we can say equivalently that $\frac{Y_n - np}{\sqrt{np(1-p)}}$ converges to standard normal in distribution.

Example 5.9 Consider rolling a die independently many times and observing $Y_n =$ the number of 6's as a function of the number of rolls n. The observed proportion is $\widehat{p}_n = \frac{Y_n}{n}$. Since $Y_n \sim$ binomial$(n, \frac{1}{6})$, we can determine that $\mathsf{E}(\widehat{p}_n) = \frac{1}{6}$ and $\mathsf{var}(\widehat{p}_n) = \frac{5/36}{n}$.

Note that $\widehat{p}_n = \frac{Y_n}{n}$ is a sample average (of iid Bernoulli trials). The SLLN tells us, therefore, that $\widehat{p}_n \to \frac{1}{6}$ w.p. 1.

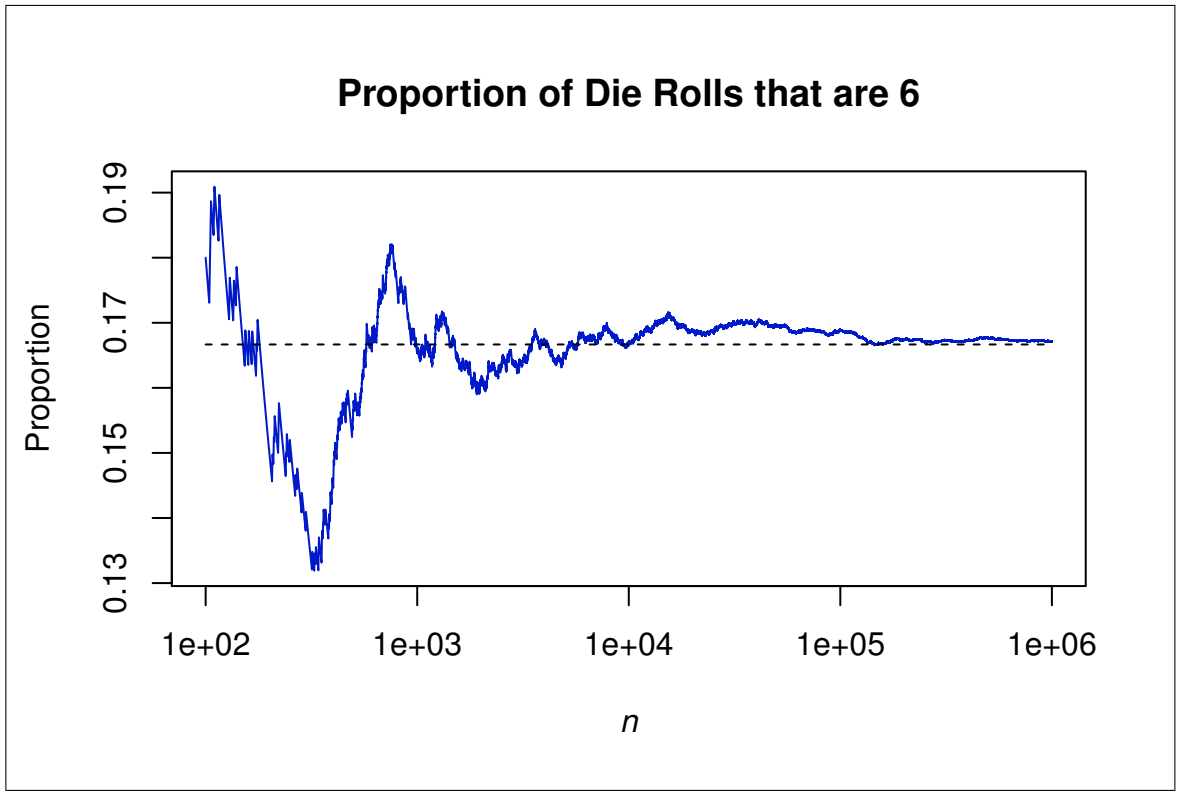

Figure 5.4 Computer simulation of proportion of 6's as sample size n increases.

On the other hand, the sequence of standardized counts

$$Z_n = \frac{Y_n - n/6}{\sqrt{n(1/6)(5/6)}} = \frac{\sqrt{n}(\widehat{p}_n - 1/6)}{\sqrt{5/36}}$$

does not itself converge, but its distribution does (to standard normal).

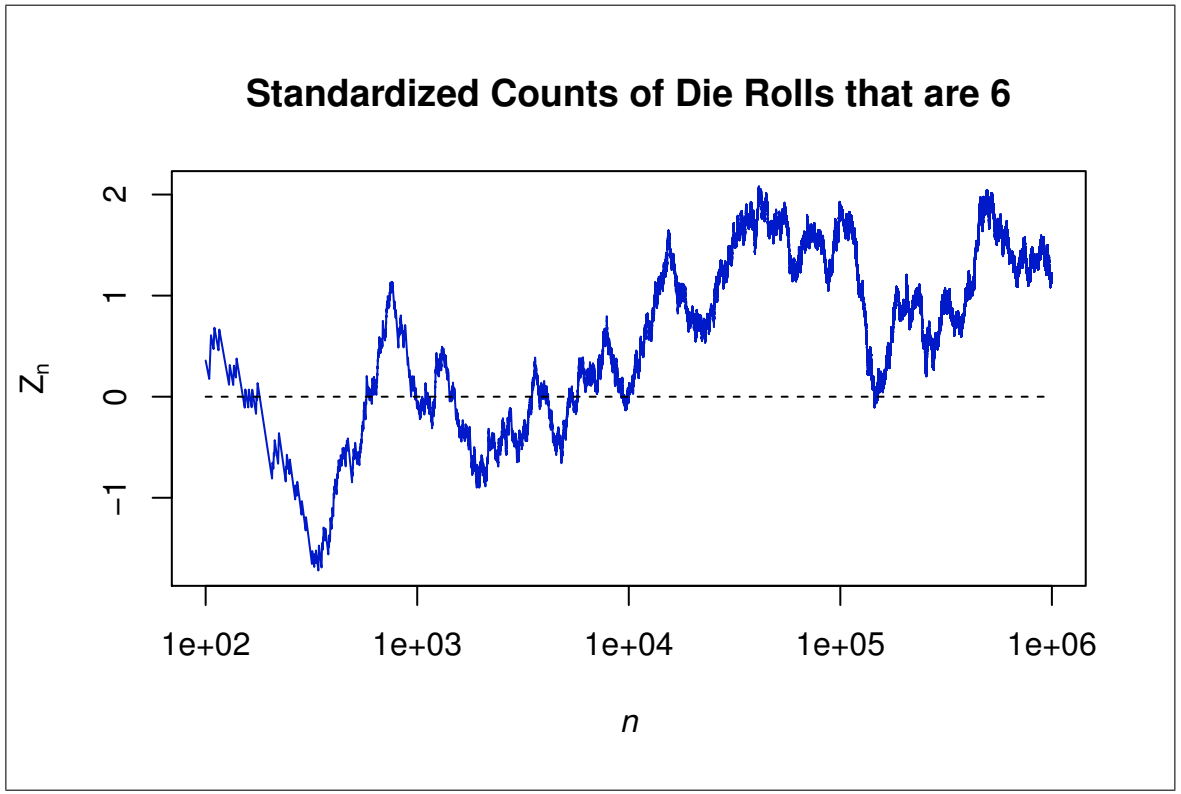

Figure 5.5 Computer simulation of standardized count of 6's as sample size n increases.

Example 5.10 Thirty independent ozone level observations of unknown distribution are taken. The observed mean is $\overline{X} = 10.47$ ppm and the observed standard deviation is $\widehat{\sigma} = 0.94$ ppm.

Suppose the EPA requires the true average to be no more than 10 ppm.

Do these data provide strong enough evidence to conclude the standard has been exceeded?

Solution Let μ and σ be the unknown mean and standard deviation of the ozone distribution. Supposing that $\mu = 10$ (worst case) and estimating σ to be 0.94, we find

$$\mathsf{P}(\overline{X} \geq 10.47) \doteq \mathsf{P}\left(\frac{\sqrt{n}(\overline{X}-10)}{\sigma} > \frac{\sqrt{30}(10.47-10)}{.94}\right) \doteq 1 - \Phi(2.73861) = .0031.$$

This is quite small, so it is very doubtful that the supposition $\mu \leq 10$ is correct.

In the last example we presumed it would be approximately correct to substitute the sample standard deviation $\widehat{\sigma}$ for σ (since $\widehat{\sigma}$ is consistent for σ).

But $\widehat{\sigma}$ is a statistic, too, and therefore random. This surely has to *affect the probability calculation*. For large enough samples, however, the effect is small, as the next theorem attests.

THEOREM 5.15 (Slutsky) Suppose Y_n converges in distribution to the cdf F_Y for some rv Y.

Suppose a and b are real numbers with $a \neq 0$ and a_n and b_n are *random sequences* such that $a_n \to a$ i.p. and $b_n \to b$ i.p.

Then $a_n Y_n + b_n \xrightarrow{\text{D}} F_{aY+b}$.

In particular, if $X_1, X_2, \ldots$ is an iid sequence with finite second moment then $T_n = \frac{\sqrt{n}(\overline{X}_n - \mu_X)}{\widehat{\sigma}_X}$ converges in distribution to the standard normal (with cdf Φ).

PROOF (Heuristic) Although the assumption of convergence in distribution does not require that the sequence Y_n converges (say, in probability), it turns out that for a proof such as this we can pretend that it does. Thus we have

$$Y_n \to Y,\ a_n \to a,\ b_n \to b \implies a_n Y_n + b_n \to aY + b,$$

and this in turn implies $a_n Y_n + b_n \xrightarrow{\text{D}} F_{aY+b}$.

To see how Slutsky's theorem applies to $\overline{X}_n$, let $Z \sim \text{normal}(0, 1)$ and $Z_n = \sqrt{n}(\overline{X}_n - \mu_X)/\sigma_X$. The CLT says Z_n converges in distribution to Φ. But we have from before that $\sigma_X/\widehat{\sigma}_X \to 1$ i.p. (and in fact w.p. 1) by Thm. 5.13.

Thus, by Slutsky's theorem, $T_n = \frac{\sigma_X}{\widehat{\sigma}_X} Z_n$ converges in distribution to Φ also. □

Example 5.2 (cont.) $Y_1, Y_2, \ldots \overset{\text{iid}}{\sim}$ Poisson(λ). Since $\mu_Y = \sigma_Y^2 = \lambda$, the CLT says $\frac{\sqrt{n}(\overline{Y}_n - \lambda)}{\sqrt{\lambda}}$ converges in distribution to the normal(0,1) dist. That is, $\overline{Y}_n$ is asymptotically normal$(\lambda, \frac{\lambda}{n})$.

It turns out that, for Poisson data, $\overline{Y}_n$ is the best estimator for λ and thus for σ_Y^2 as well.

As it is also true that $\lambda/\overline{Y}_n \to 1$ w.p. 1, then we have $\frac{\sqrt{n}(\overline{Y}_n - \lambda)}{\sqrt{\overline{Y}_n}}$ converges in distribution to the normal(0,1) distribution, by Slutsky's theorem.

(However, there is a slight technical problem with that because the probability that $\overline{Y}_n = 0$ is positive, albeit extremely small, no matter how large n. This implies that the random variable $\frac{\sqrt{n}(\overline{Y}_n - \lambda)}{\sqrt{\overline{Y}_n}}$ actually has an *infinite mean*, despite the convergence in distribution.)

The examples above all are about the asymptotic distribution of an average.

Often, however, we are more interested in some (nice) *function of an average*.

While we can handle that using the basic CLT and the methods of transformations of random variables in Chapter 2, it sometimes is more convenient to extend the CLT as follows.

THEOREM 5.16 (Delta Method) Let $X_1, X_2, \ldots$ be iid as in Thm. 5.14.

If $g(x)$ is a *continuously differentiable* function then

$$\sqrt{n}(g(\overline{X}_n) - g(\mu_X)) \xrightarrow{\text{D}} \text{normal}(0, (g'(\mu_X))^2\sigma_X^2) \text{ in distribution.}$$

**** Combined with Slutsky's method, the delta method enables us to find useful inference methods not just for expectations but also for functions of expectations.*

PROOF (Heuristic) By Taylor's theorem,

$$g(\overline{X}_n) - g(\mu_X) = g'(\mu_X)(\overline{X}_n - \mu_X) + \text{ lower order vanishing terms.}$$

Since $\sqrt{n}(\overline{X}_n - \mu_X) \xrightarrow{\text{D}} \text{normal}(0, \sigma_X^2)$ by the CLT, it follows that $\sqrt{n}(g(\overline{X}_n) - g(\mu_X)) \xrightarrow{\text{D}} \text{normal}(0, (g'(\mu_X))^2\sigma_X^2)$. □

Example 5.1 (cont.) Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ exponential(β).

Since $\mathsf{E}(X_i) = \beta$, $\overline{X}_n$ is a natural estimator for β. (It turns out that it is also the optimal estimator.)

Also, $\mathsf{var}(X_i) = \beta^2$. By the CLT we know that $\sqrt{n}(\overline{X}_n - \beta) \xrightarrow{\text{D}}$ normal$(0, \beta^2)$.

However, one is also often interested in the so-called *rate parameter* $g(\beta) = \frac{1}{\beta}$ which is estimated by $\frac{1}{\overline{X}_n}$.

Note that $g'(\beta) = -\frac{1}{\beta^2}$. The delta method then gives

$$\sqrt{n}\left(\frac{1}{\overline{X}_n} - \frac{1}{\beta}\right) \xrightarrow{\text{D}} \text{normal}\left(0, \left(\frac{1}{\beta^2}\right)^2 \beta^2\right).$$

That is, $1/\overline{X}_n$ is asymptotically normal$(\frac{1}{\beta}, \frac{1}{n\beta^2})$.

Applying Slutksy as well, we also have

$$\sqrt{n}\,\overline{X}_n\left(\frac{1}{\overline{X}_n} - \frac{1}{\beta}\right) \xrightarrow{\text{D}} \text{normal}(0, 1).$$

Example 5.8 (cont.) Consider an iid sequence of Bernoulli(p) trials.

The *odds* $g(p) = \frac{p}{1-p}$ is often a quantity of interest. Note that $g'(p) = \frac{1}{(1-p)^2}$.

The sample proportion of successes $\widehat{p}_n$ is asymptotically normal$(p, \frac{p(1-p)}{n})$.

Then the delta method says that the sampling distribution for the sample odds $g(\widehat{p}_n) = \frac{\widehat{p}_n}{1-\widehat{p}_n}$ is approximately normal$(\frac{p}{1-p}, \frac{p}{n(1-p)^3})$.

We might also be interested in the *log-odds* $h(p) = \log(\frac{p}{1-p})$. Since $h'(p) = \frac{1}{p(1-p)}$, it follows that

$$\sqrt{n}\left(\log\left(\frac{\widehat{p}_n}{1-\widehat{p}_n}\right) - \log\left(\frac{p}{1-p}\right)\right) \xrightarrow{\text{D}} \text{normal}\left(0, \frac{1}{p(1-p)}\right).$$

And by Slutsky, we can also say that

$$\sqrt{n\widehat{p}_n(1-\widehat{p}_n)}\left(\log\left(\frac{\widehat{p}_n}{1-\widehat{p}_n}\right) - \log\left(\frac{p}{1-p}\right)\right) \xrightarrow{\text{D}} \text{normal}(0, 1).$$

Example 5.11 The Maxwell(β) distribution has pdf

$$f_W(w) = \sqrt{\frac{2}{\pi}}\frac{w^2}{\beta^3}\mathrm{e}^{-w^2/(2\beta^2)},\ w > 0.$$

Note that this forms an exponential family with $t(w) = w^2$.

Recall Thm. 3.18, which describes how certain expectations can be found by taking derivatives with respect to the parameter(s). We compute

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}\beta}\int_0^\infty f_W(w)\,\mathrm{d}w = \int_0^\infty \left(\frac{-3}{\beta} + \frac{w^2}{\beta^3}\right)\sqrt{\frac{2}{\pi}}\frac{w^2}{\beta^3}\mathrm{e}^{-w^2/(2\beta^2)}\,\mathrm{d}w \\ &= \frac{-3}{\beta} + \frac{\mathsf{E}(W^2)}{\beta^3}. \end{aligned}$$

From this we get $\mathsf{E}(W^2) = 3\beta^2$. (Alternatively, show that $Y = W^2$ has a gamma distribution and use that.)

Similarly,

$$0 = \frac{\mathrm{d}^2}{\mathrm{d}^2\beta}\int_0^\infty f_W(w)\,\mathrm{d}w = \frac{6}{\beta^2} - \frac{9\mathsf{E}(W^2)}{\beta^4} + \frac{\mathsf{E}(W^4)}{\beta^6},$$

yielding $\mathsf{var}(W^2) = \mathsf{E}(W^4) - (\mathsf{E}(W^2))^2 = 12\beta^4$.

Now imagine an iid sequence $W_1, W_2, \ldots$ from the Maxwell(β) distribution. Then

$$\widehat{\beta^2} = \frac{1}{3n}\sum_{i=1}^{n} W_i^2 \to \frac{1}{3}\mathsf{E}(W^2) = \beta^2$$

by the SLLN. Also, by the CLT and the var(W^2) computed above,

$$\sqrt{n}(\widehat{\beta^2} - \beta^2) \xrightarrow{\mathrm{D}} \text{normal}(0, 4\beta^4/3).$$

However, the estimator for β is $\widehat{\beta} = \sqrt{\frac{1}{3n}\sum_{i=1}^{n} W_i^2}$. Here we can again use the delta method, this time with $g(x) = \sqrt{x}$ and $g'(x) = \frac{1}{2\sqrt{x}}$. Thus,

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{\mathrm{D}} \text{normal}(0, (g'(\beta^2))^2(4\beta^4/3)) = \text{normal}(0, \beta^2/3).$$

5.4 Random Normal Samples

The classic theory for sampling distributions was developed for the situation that the data have *normal* distribution. Even for non-normal data, this material inspires and motivates much theory and methodology.

So, even though it is a very special assumption, we take the time to cover several important consequences of having data from the normal distribution.

We start with describing the exact *joint sampling distribution* of the sample mean and variance, $(\overline{X}, S^2)$.

THEOREM 5.17 Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{normal}(\mu, \sigma^2)$.

i. $\overline{X} \sim \text{normal}(\mu, \sigma^2/n)$.

ii. $\frac{(n-1)S^2}{\sigma^2} = \frac{n\widehat{\sigma}^2}{\sigma^2} \sim \text{chi-square}(n-1)$. Equivalently, $S^2 \sim \text{gamma}(\frac{n-1}{2}, \frac{2\sigma^2}{n-1})$.

iii. $\overline{X}$ and S^2 are independent.

PROOF (sketch) $\frac{(n-1)S^2}{\sigma^2}$ can be expressed as $Z_1^2 + \cdots + Z_{n-1}^2$ where

$$Z_i = \frac{\sum_{j=1}^{i} X_j - iX_{i+1}}{\sqrt{i(i+1)}\sigma}.$$

With a lot of algebra it may be shown that each Z_i is a linear combination of $X_1, \ldots, X_n$ that has mean 0, variance 1 and is *uncorrelated* with $\overline{X}$ and all the other Z_i's.

By our bivariate (and an analogous multivariate) normal theory, we find that $Z_1, \ldots, Z_{n-1}$ and $Z_n = \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}$ are *independent* normal(0,1) rvs.

It follows that $\overline{X}$ is independent of S^2 and, by Thm. 3.11, $Z_1^2, \ldots, Z_{n-1}^2$ are iid chi-square(1) rvs.

Since chi-square(m) is the same as gamma($m/2, 2$) then the sum of $n-1$ chi-square(1) independent rvs has chi-square($n-1$) distribution (recall Ex. 5.5), and this completes the argument. □

The significance of this result, stated next, is that with *normal data* we do not have to rely on limit theory to approximate probabilities about $\overline{X}_n$ when σ must be estimated.

THEOREM 5.18 (Gosset, a.k.a. "Student") Suppose $Z \sim \text{normal}(0,1)$ and $Y \sim \text{chi-square}(m)$ are independent rvs.

Then $T = \sqrt{m}Z/\sqrt{Y}$ has Student's t-distribution with pdf

$$f_T(t) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\sqrt{m\pi}}\left(\frac{1}{1+t^2/m}\right)^{(m+1)/2}.$$

The parameter m is called the degrees of freedom.

In particular, suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{normal}(\mu, \sigma^2)$. Then $T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S}$ has Student's t-distribution with $n-1$ degrees of freedom.

PROOF Let $V = Y + Z^2$. Then $1 + t^2/m = 1 + z^2/y = v/y$,

$$\frac{\mathrm{d}t\mathrm{d}v}{\mathrm{d}z\mathrm{d}y} = \left|\det\begin{pmatrix} \frac{\sqrt{m}}{\sqrt{y}} & \frac{\sqrt{m}z}{2y^{3/2}} \\ 2z & 1 \end{pmatrix}\right| = \frac{\sqrt{m}}{\sqrt{y}}(1 + z^2/y)$$

and

$$\begin{aligned} f_{T,V}(t,v) &= \text{const.} \times \mathrm{e}^{-z^2/2} y^{\frac{m}{2}-1} \mathrm{e}^{-y/2} \tfrac{y^{1/2}}{1+z^2/y} \\ &= \text{const.} \times \mathrm{e}^{-v/2} \left(\frac{v}{1+t^2/m}\right)^{(m-1)/2} \frac{1}{1+t^2/m} \\ &= \text{const.} \times v^{\frac{m+1}{2}-1} \mathrm{e}^{-v/2} \times \left(\frac{1}{1+t^2/m}\right)^{(m+1)/2}. \end{aligned}$$

This factors into the *kernel* for the t-distribution with m degrees of freedom (namely, $(1+t^2/m)^{-(m+1)/2}$) and the *kernel* for the chi-square$(m+1)$ distribution (namely, $v^{\frac{m+1}{2}-1}e^{-v/2}$).

Consequently, this shows that T has the required distribution and that T is independent of V as well.

The constant is obtained with a little more care, or by noting that T^2/m is the rv R in Ex. 4.12 (and $\alpha = 1/2$, $\beta = m/2$).

The distribution of $T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S}$ follows from Thm. 5.17 and the above with $Z = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$, $Y = \frac{(n-1)S^2}{\sigma^2}$ and $m = n - 1$. □

- The application for this theorem is the same as Ex. 5.10 except that we can compute the probability *exactly* from Student's t-distribution. There is no need to appeal to the CLT and Slutsky's theorem, but we do *assume* the data have a normal distribution.
- On the other hand, statisticians often will use the t-distribution anyway because it typically results in a more *conservative approach* to statistical inference and therefore may overcome any inaccuracies due to the CLT approximation.
- Unfortunately, the independence of $\overline{X}$ and S^2 is crucial. They are at least uncorrelated when the data come from a symmetric distribution. In this case the approximation is generally reasonable.
- But (as seen in Thm. 5.8) the correlation of $\overline{X}$ and S^2 is a function of the skewness coefficient. Consequently, using the t-distribution for inference can be quite inaccurate when the data come from a very *asymmetric* distribution, at least if n is not very large.

Example 5.10 (cont.) 30 ozone levels with $\overline{X} = 10.47$ and $\widehat{\sigma} = .94$. Assume the ozone levels are normal and recompute the probability from earlier, using a t-distribution with $n - 1 = 29$ degrees of freedom.

Solution First, $S = \sqrt{\frac{n}{n-1}}\,\widehat{\sigma} = 0.95607$.

Then we calculate (with $\mu = 10$, as before)

$$\mathsf{P}(\overline{X} \geq 10.47) = \mathsf{P}(\tfrac{\sqrt{n}(\overline{X}-\mu)}{S} > 2.6926) = \mathsf{P}(T_n > 2.6926) = .0058,$$

which is somewhat larger than the normal approximation we used before.

(Note, $T_n = \frac{\sqrt{n-1}(\overline{X}-\mu)}{\widehat{\sigma}}$, when expressed in terms of $\widehat{\sigma}$ instead of S.)

Even if the data are not normal, using a t-distribution (as opposed to the normal distribution) generally will be more accurate for calculations such as the above, if σ must be estimated.

On the other hand, both methods can be somewhat inaccurate if the data are sampled from a very asymmetric distribution.

Ex. 5.10 is an example of hypothesis testing: determining whether a specific value is *plausible* for a parameter, in light of the data.

Actual estimation of the parameter with bounds for error is also very important. Thm. 5.18 gives us the quintessential method for doing this.

Example 5.3 (cont.) Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{normal}(\mu, \sigma^2)$.

Let $\alpha > 0$ (small) and find *statistics* L and U (lower and upper bounds) such that $\text{P}(L \leq \mu \leq U) = 1 - \alpha$.

Solution By Thm. 5.18, $T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S}$ has t-distribution with $n-1$ degrees of freedom. Note that this distribution is symmetric about 0.

Choose a constant $t_{\alpha/2}$ such that $\text{P}(T_n < -t_{\alpha/2}) = \text{P}(T_n > t_{\alpha/2}) = \alpha/2$. Then, "solving" for μ,

$$\begin{aligned} 1 - \alpha &= \text{P}\left(-t_{\alpha/2} \leq \frac{\sqrt{n}(\overline{X}_n - \mu)}{S} \leq t_{\alpha/2}\right) \\ &= \text{P}\left(\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right). \end{aligned}$$

Thus we take $L = \overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}$ and $U = \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}$.

These are *statistics* because they do not depend on either μ or σ^2 to be calculated.

The *random* interval $[L, U]$ "captures" the (unknown) value of μ with probability $1 - \alpha$ and it therefore is called a $1 - \alpha$ confidence interval for μ.

- By Thm. 5.15 we know that $T_n \to \text{normal}(0,1)$ in distribution. Thus, with large enough degrees of freedom, probabilities for the t-distribution can be approximated accordingly.
- The probability tails for the t-distribution, however, damp as a power function, not exponentially like they do for the normal cdf. So the normal approximation is not always real close, at least relatively, for tail probabilities. (Which, unfortunately, are the cases statisticians are most interested in.)
- The t-distribution will often be used any time the data are approximately normal and/or the sample size is large.
- It can even be used when one has an estimator that has approximately normal distribution (such as a sample moment) and its variance *must be estimated*. Thus, for example for the second sample moment,

$$\frac{\sqrt{n-1}(\widehat{\mu}_2' - \mu_2')}{\sqrt{\widehat{\mu}_4' - (\widehat{\mu}_2')^2}}$$

will have approximately a t-distribution with $n-1$ degrees of freedom, at least for large n.

Assuming normality leads to many nice results about the distributions of estimators. As an example, we mention the problem of comparing two variances.

THEOREM 5.19 Suppose $X_1, \ldots, X_m \overset{\text{iid}}{\sim} \text{normal}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{normal}(\mu_Y, \sigma_Y^2)$ and the two samples are independent.

Let S_X^2 and S_Y^2 be the respective variance estimates. Then $\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$ has F-distribution (see Ex. 4.12) with parameters (degrees of freedom) $m-1$ and $n-1$.

PROOF By Thm. 5.17.ii, $(m-1)S_X^2/\sigma_X^2$ and $(n-1)S_Y^2/\sigma_Y^2$ are each chi-square random variables, and they are independent. The degrees of freedom are $m-1$ and $n-1$, respectively.

We found the pdf for the ratio of independent gamma random variables in Ex. 4.12 and provided its pdf. Translating that to the context here yields the pdf of the F-distribution. □

With this result it is possible to discuss the sampling distribution of the estimator for a ratio of sample variances such as S_X^2/S_Y^2.

5.5 Vector-Valued Statistics

The purpose of this section is to consider how to handle asymptotics for a vector of statistics or estimators. To do that, we also need to give brief consideration to looking at multivariate data, with an emphasis on the multivariate normal model.

Suppose that $(\underline{X}_1, \ldots, \underline{X}_n)$ is a simple random sample of random *k-dimensional vectors* from some joint distribution F, where we represent the observations as $\underline{X}_m = (X_{m,1}, \ldots, X_{m,k})$. Consider then the sample mean vector

$$\overline{\underline{X}} = (\overline{X}_1, \ldots, \overline{X}_k) = \left(\frac{1}{n}\sum_{m=1}^{n} X_{m,1}, \ldots, \frac{1}{n}\sum_{m=1}^{n} X_{m,k}\right).$$

- Clearly, $\mathsf{E}(\overline{X}_j) = \mathsf{E}(X_{1,j})$ for each $j = 1, \ldots, k$, and $\overline{X}_j \to \mathsf{E}(X_{1,j})$ as $n \to \infty$ by the SLLN.
- Moreover, each $\overline{X}_j$ is asymptotically normal, as $\sqrt{n}(\overline{X}_j - \mathsf{E}(X_{1,j})) \xrightarrow{\mathrm{D}}$ normal$(0, \mathsf{var}(X_{1,j}))$.
- Our objective is to describe a similar distributional limit *jointly* for the vector $\overline{\underline{X}}$.

First, we extend Def. 4.37 (the *bivariate* normal distribution). However, our definition will be in terms that we can use in a proof, rather than by giving the joint pdf (which is best expressed using matrix notation).

DEFINITION 5.20 A random vector $\underline{X} = (X_1, \ldots, X_k)$ has multivariate normal distribution if $Y = a_1X_1 + \cdots + a_kX_k$ has normal distribution for any real-valued constants $a_1, \ldots, a_k$.

The multivariate normal distribution is characterized by its means $\mu_i = \mathsf{E}(X_i)$ and its covariances $\sigma_{i,j} = \mathsf{cov}(X_i, X_j)$. ($\sigma_{i,i} = \mathsf{var}(X_i)$.)

- The means are often denoted with a vector $\underline{\mu} = (\mu_1, \ldots, \mu_k)$ and the covariances with a matrix Σ having element $\sigma_{i,j}$ in the i-th row and j-th column. Note that Σ is *symmetric* since $\sigma_{i,j} = \sigma_{j,i}$.
- Clearly, the definition implies that the marginal distributions are normal, and each *sub-vector* has a multivariate normal distribution.
- We have $\mathsf{E}(a_1X_1 + \cdots + a_kX_k) = \sum_{i=1}^k a_i\mu_i$ and (extending Thm. 4.34.ii) $\mathsf{var}(a_1X_1 + \cdots + a_kX_k) = \sum_{i=1}^k \sum_{j=1}^k a_ia_j\sigma_{i,j}$. These are generally true.

Now we provide the primary result.

THEOREM 5.21 (Multivariate CLT) Suppose, as previously, that $(\underline{X}_1, \ldots, \underline{X}_n)$ is an iid sample of random k-dimensional vectors and let $\overline{\underline{X}} = (\overline{X}_1, \ldots, \overline{X}_k)$ be the vector of sample means.

Also define $\underline{\mu}$ to be the vector with elements $\mu_i = \mathsf{E}(X_{1,i})$ and Σ to be the matrix with elements $\sigma_{i,j} = \mathsf{cov}(X_{1,i}, X_{1,j})$ (all assumed to be finite).

Then

$$\sqrt{n}(\overline{\underline{X}} - \underline{\mu}) \xrightarrow{\mathrm{D}} \text{mult. normal}(\underline{0}, \Sigma), \quad \text{as } n \to \infty.$$

PROOF (Sketch) Fix arbitrary real-valued constants $a_1, \ldots, a_k$ and define $Y_m = a_1 X_{m,1} + \cdots + a_k X_{m,k}$. Obviously $\overline{Y} = a_1\overline{X}_1 + \cdots + a_k\overline{X}_k$, which has expectation, and converges to, $\mathsf{E}(Y_1) = a_1\mu_1 + \cdots + a_k\mu_k$. Additionally, (using the comments above),

$$\mathsf{var}(\overline{Y}) = \frac{\mathsf{var}(Y_i)}{n} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j \sigma_{i,j}.$$

Moreover, $\sqrt{n}(\overline{Y} - \mathsf{E}(Y_m)) \xrightarrow{\mathrm{D}} \text{normal}(0, \mathsf{var}(Y_m))$. Since the a_i's are arbitrary, this suffices to show joint convergence to the multivariate normal distribution. □

Example 5.12 Suppose (X_m, Y_m), $m = 1, \ldots, n$ are iid random vectors with joint pdf

$$f_{X,Y}(x, y) = 4x(x^2 + y - 2x^2y)1_{(0,1)}(x)1_{(0,1)}(y).$$

It is easy to find $\mathsf{E}(X) = \frac{1}{3}$, $\mathsf{E}(Y) = \frac{1}{2}$, and

$$\mathsf{var}(X) = \frac{5}{36}, \quad \mathsf{var}(Y) = \frac{1}{12}, \quad \mathsf{cov}(X, Y) = \frac{13}{90}.$$

Then $\sqrt{n}(\overline{X} - \frac{1}{3}, \overline{Y} - \frac{1}{2}) \xrightarrow{\mathrm{D}}$ biv. normal$((0, 0), \Sigma)$, where $\Sigma = \begin{pmatrix} \frac{5}{36} & \frac{13}{90} \\ \frac{13}{90} & \frac{1}{12} \end{pmatrix}$.

Example 5.13 (Multinomial) Consider a random sample $C_1, \ldots, C_n$ where each C_m is a *categorial* rv taking one of $k \geq 3$ possible values (say, $1, 2, \ldots, k$) with respective positive probabilities $p_1, p_2, \ldots, p_k$. $(p_1 + \cdots + p_k = 1.)$

Let $X_{m,i} = 1_{\{i\}}(C_m)$ be the indicator that observation m is in category i. Then $Y_i = \sum_{m=1}^{n} X_{m,i}$ is a sum of iid Bernoulli rvs and has binomial(n, p_i) distribution.

The vector $(Y_1, \ldots, Y_k)$ has multinomial$(n, p_1, \ldots, p_k)$ distribution. Using principles of counting and independence of the C_i's, the joint pmf is

$$f(y_1, \ldots, y_k) = \frac{n!}{y_1! \cdots y_k!} p_1^{y_1} \cdots p_k^{y_k}, \quad y_1 + \cdots + y_k = n.$$

(This generalizes Ex. 4.1 from Chapter 4.)

We know that, since Y_i has binomial distribution, $\mathsf{E}(Y_i) = np_i$ and $\mathsf{var}(Y_i) = np_i(1-p_i)$. Moreover, by the basic CLT,

$$\frac{Y_i - np_i}{\sqrt{n}} \xrightarrow{\text{D}} \mathsf{normal}(0, p_i(1-p_i)).$$

Let $\sigma_{i,i} = p_i(1-p_i)$.

Now we observe that, if $i \neq j$ then $X_{m,i}X_{m,j} = 0$ since C_m cannot be both i and j. Hence,

$$\sigma_{i,j} = \mathsf{cov}(X_{m,i}, X_{m,j}) = 0 - \mathsf{E}(X_{m,i})\mathsf{E}(X_{m,j}) = -p_ip_j \quad \text{if } i \neq j,$$

and $\mathsf{cov}(Y_i, Y_j) = -np_ip_j$ if $i \neq j$.

At this point, we observe that $Y_1, \ldots, Y_k$ are *linearly dependent*, and each $\sum_{i=1}^k \sigma_{i,j} = 0$. So to get a proper joint limit distribution we will reduce the vector by one component to $(Y_1, \ldots, Y_{k-1})$. Let Σ denote the matrix of covariance $\sigma_{i,j}$ with $1 \leq i \leq k-1$ and $1 \leq j \leq k-1$.

By the multivariate CLT, it follows that

$$\left(\frac{Y_1 - np_1}{\sqrt{n}}, \ldots, \frac{Y_{k-1} - np_{k-1}}{\sqrt{n}}\right) \xrightarrow{\text{D}} \mathsf{mult.\ normal}(0, \Sigma).$$

However, the application of Thm. 5.21 need not be just for multivariate random samples.

Suppose $X_1, X_2, \ldots,$ are iid random variables with finite $2k$-moment and let $\mu'_i = \mathsf{E}(X^i)$ be the i-th moment for $i \leq 2k$.

Recall the sample i-th moment $\widehat{\mu}'_i = \frac{1}{n}\sum_{m=1}^{n} X_m^i$, which is unbiased for μ'_i. Since $i \leq 2k$,

$$\mathsf{var}(\widehat{\mu}'_i) = \frac{1}{n}\mathsf{var}(X^i) = \mathsf{E}((X^i)^2) - (\mathsf{E}(X^i))^2 = \frac{1}{n}(\mu'_{2i} - (\mu'_i)^2).$$

Each $\widehat{\mu}'_i$ is asymptotically normal, by the ordinary CLT.

But now we want asymptotics for the *vector* of sample moments, $\underline{\widehat{\mu}'} = (\widehat{\mu}'_1, \ldots, \widehat{\mu}'_k)$.

Applying Thm. 5.21, we have this result on the following slide.

COROLLARY 5.22 Assume as above. Define $\underline{\mu}'$ to be the vector of the first k moments and $\underline{\widehat{\mu}}'$ to be the vector of the first k sample moments. Also, define Σ to be the matrix with elements

$$\sigma_{i,j} = \mu'_{i+j} - \mu'_i \mu'_j \quad 1 \le i \le k, 1 \le j \le k.$$

Then $\sqrt{n}(\underline{\widehat{\mu}}' - \underline{\mu}') \xrightarrow{\text{D}}$ mult. normal$(\underline{0}, \Sigma)$.

PROOF Consider the random vectors $\underline{X}_i = (X_i, X_i^2, \ldots, X_i^k)$. The vector of means is $\underline{\widehat{\mu}}' = (\widehat{\mu}'_1, \ldots, \widehat{\mu}'_k)$.

We merely need to observe that

$$\text{cov}(X^i, X^j) = \text{E}(X^{i+j}) - \text{E}(X^i)\text{E}(X^j) = \mu'_{i+j} - \mu'_i \mu'_j,$$

and the rest follows from the multivariate CLT.

Note that $\text{cov}(\widehat{\mu}'_i, \widehat{\mu}'_j) = \frac{\mu'_{i+j} - \mu'_i \mu'_j}{n}$. This includes the case $i = j$. □

If needed for statistical purposes, $\sigma_{i,j}$ can be *estimated* with $\widehat{\sigma}_{i,j} = \widehat{\mu}'_{i+j} - \widehat{\mu}'_i \widehat{\mu}'_j$.

Example 5.14 Let $T_i \overset{\text{iid}}{\sim} \text{gamma}(\alpha, \beta)$. Then $\widehat{\mu}_1' = \overline{T}$ has mean $\mu = \mu_1' = \alpha\beta$ and variance $\frac{\alpha\beta^2}{n}$, since $\sigma^2 = \alpha\beta^2$.

Additionally, $\widehat{\mu}_2' = \frac{1}{n}\sum_{i=1}^n T_i^2$ has mean $\mu_2' = \alpha(\alpha+1)\beta^2$ and variance

$$\frac{\mu_4' - (\mu_2')^2}{n} = \ldots = \frac{2\alpha(\alpha+1)(2\alpha+3)\beta^4}{n}.$$

We can also find

$$\text{cov}(\widehat{\mu}_1', \widehat{\mu}_2') = \frac{\mu_3' - \mu_1'\mu_2'}{n} = \ldots = \frac{2\alpha(\alpha+1)\beta^3}{n}.$$

And then $(\widehat{\mu}_1', \widehat{\mu}_2')$ is accordingly asymptotic bivariate normal by Cor. 5.22.

We would also like to determine the asymptotic nature of the sample variance $\widehat{\sigma}^2 = \widehat{\mu}_2' - (\widehat{\mu}_1')^2$ and of the sample standard deviation $\widehat{\sigma}$. Better still, we would like to get an asymptotic joint distribution for $(\widehat{\mu}_1', \widehat{\sigma}^2)$.

However, $\widehat{\sigma}^2$ is *not a linear transformation* of the ordinary (uncentered) sample moments. Hence, we need yet another extension theorem – this time for the *delta method*.

Rather than give a general theorem, we will just consider the problem described on the previous slide. The proof is sufficiently illustrative to suggest how to extend the method.

THEOREM 5.23 Let $\overline{X}$ and $\widehat{\sigma}^2$ be the sample mean and variance, respectively, from a random sample from a distribution with mean μ and variance σ^2. Assume $k = 4$ finite moments and define $\sigma_{i,j}$ as in Cor. 5.22.

Then

$$\sqrt{n}\begin{pmatrix}\overline{X} - \mu \\ \widehat{\sigma}^2 - \sigma^2\end{pmatrix} \xrightarrow{\text{D}} \text{biv. normal}((0,0), \Sigma^*),$$

where

$$\Sigma^* = \begin{pmatrix}\sigma_{1,1} & -2\mu\sigma_{1,1} + \sigma_{1,2} \\ -2\mu\sigma_{1,1} + \sigma_{1,2} & 4\mu^2\sigma_{1,1} - 4\mu\sigma_{1,2} + \sigma_{2,2}\end{pmatrix}.$$

(Note: $\sigma_{1,1} = \sigma^2$.)

PROOF (Heuristic) We will *linearize* $\widehat{\sigma}^2$ (delta method) and derive the asymptotic normal distribution accordingly.

Let $g_1(m_1, m_2) = m_1$ and $g_2(m_1, m_2) = m_2 - m_1^2$. Then

$$\mu = g_1(\mu, \mu_2') \quad \text{and} \quad \sigma^2 = g_2(\mu, \mu_2').$$

Since $\overline{X} \to \mu$ and $\widehat{\mu}_2' \to \mu_2'$, a first-order Taylor's expansion gives

$$\begin{aligned}\widehat{\sigma}^2 - \sigma^2 &= g_2(\overline{X}, \widehat{\mu}_2') - g_2(\mu, \mu_2') \\ &\approx \left(\frac{\partial g_2}{\partial m_1}\Big|_{m_1=\mu, m_2=\mu_2'} (\overline{X} - \mu) + \frac{\partial g_2}{\partial m_2}\Big|_{m_1=\mu, m_2=\mu_2'} (\widehat{\mu}_2' - \mu_2')\right) \\ &= -2\mu(\overline{X} - \mu) + (\widehat{\mu}_2' - \mu_2').\end{aligned}$$

**** The expansion only works because $\overline{X} \to \mu$ and $\widehat{\mu}_2' \to \mu_2'$ in probability.*

We could do a similar Taylor's expansion for g_1 except it is trivial in this case.

The uptake is that $\widehat{\sigma}^2$ is approximately a linear combination of the first two sample moments. We can then get an approximate variance by

$$\text{var}(\widehat{\sigma}^2) \approx \text{var}(-2\mu\overline{X} + \widehat{\mu}_2') = \frac{(-2\mu)^2\sigma_{1,1} + 2(-2\mu)\sigma_{1,2} + \sigma_{2,2}}{n}.$$

Similarly,

$$\text{cov}(\overline{X}, \widehat{\sigma}^2) \approx \text{cov}(\overline{X}, -2\mu\overline{X} + \widehat{\mu}_2') = \frac{-2\mu\sigma_{1,1} + \sigma_{1,2}}{n} = \frac{\text{E}((X-\mu)^3)}{n}.$$

Moreover, the approximate linearity suffices for $\sqrt{n}(\overline{X} - \mu, \widehat{\sigma}^2 - \sigma^2)$ to converge in distribution to a bivariate normal, with the given covariance matrix. □

Example 5.14 (cont.) We have iid $X_i \sim \text{gamma}(\alpha, \beta)$. From earlier, the limit normal distribution for $\sqrt{n}((\widehat{\mu}'_1, \widehat{\mu}'_2) - (\mu'_1, \mu'_2))$ has covariance matrix

$$\Sigma = \begin{pmatrix} \alpha\beta^2 & 2\alpha(\alpha+1)\beta^3 \\ 2\alpha(\alpha+1)\beta^3 & 2\alpha(\alpha+1)(2\alpha+3)\beta^4 \end{pmatrix}.$$

Note that the mean is $\mu = \mu'_1$ and the sample mean is $\widehat{\mu} = \widehat{\mu}'_1$. We will simplify expressions accordingly.

By Thm. 5.23, the limit normal distribution for $\sqrt{n}((\widehat{\mu},\widehat{\sigma}^2) - (\mu,\sigma^2))$ has covariance matrix

$$\begin{aligned} \Sigma^* &= \begin{pmatrix} \sigma_{1,1} & -2\mu'_1\sigma_{1,1} + \sigma_{1,2} \\ -2\mu'_1\sigma_{1,1} + \sigma_{1,2} & 4(\mu'_1)^2\sigma_{1,1} - 4\mu'_1\sigma_{1,2} + \sigma_{2,2} \end{pmatrix} \\ &= \begin{pmatrix} \alpha\beta^2 & 2\alpha\beta^3 \\ 2\alpha\beta^3 & 2\alpha(\alpha+3)\beta^4 \end{pmatrix}. \end{aligned}$$

Now consider that

$$\widehat{\alpha} \stackrel{\text{def}}{=} \frac{\widehat{\mu}^2}{\widehat{\sigma}^2} \to \frac{\mu^2}{\sigma^2} = \alpha \quad \text{and} \quad \widehat{\beta} \stackrel{\text{def}}{=} \frac{\widehat{\sigma}^2}{\widehat{\mu}} \to \frac{\sigma^2}{\mu} = \beta.$$

What can we say about the limit distribution for $\sqrt{n}((\widehat{\alpha}, \widehat{\beta}) - (\alpha, \beta))$?

To do this, we have to master the *full multivariate version* of the delta method. We will not prove it here, but simply take inspiration from the proof of Thm. 5.23.

To this end, let $h_1(\mu, \sigma^2) = \frac{\mu^2}{\sigma^2}$ and $h_2(\mu, \sigma^2) = \frac{\sigma^2}{\mu}$. Then obtain the *matrix of partial derivatives*

$$D = \begin{pmatrix} \frac{\partial h_1}{\partial \mu} & \frac{\partial h_1}{\partial \sigma^2} \\ \frac{\partial h_2}{\partial \mu} & \frac{\partial h_2}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{2\mu}{\sigma^2} & -\frac{\mu^2}{\sigma^4} \\ -\frac{\sigma^2}{\mu^2} & \frac{1}{\mu} \end{pmatrix} = \begin{pmatrix} \frac{2}{\beta} & -\frac{1}{\beta^2} \\ -\frac{1}{\alpha} & \frac{1}{\alpha\beta} \end{pmatrix}.$$

The delta method approximates the *nonlinear* transformations of the statistics as *linear* transformations using these partial derivatives, and then it finds the variances and covariances of those linear transformations.

This leads to the new covariance matrix

$$\begin{aligned} \Sigma^{**} &= D\Sigma^* D^T \\ &= \begin{pmatrix} \frac{2}{\beta} & -\frac{1}{\beta^2} \\ -\frac{1}{\alpha} & \frac{1}{\alpha\beta} \end{pmatrix} \begin{pmatrix} \alpha\beta^2 & 2\alpha\beta^3 \\ 2\alpha\beta^3 & 2\alpha(\alpha+3)\beta^4 \end{pmatrix} \begin{pmatrix} \frac{2}{\beta} & -\frac{1}{\alpha} \\ -\frac{1}{\beta^2} & \frac{1}{\alpha\beta} \end{pmatrix} \\ &= \begin{pmatrix} 2\alpha(\alpha+1) & -2(\alpha+1)\beta \\ -2(\alpha+1)\beta & \frac{2\alpha+3}{\alpha}\beta^2 \end{pmatrix}. \end{aligned}$$

We conclude that $\sqrt{n}((\widehat{\alpha}, \widehat{\beta}) - (\alpha, \beta)) \xrightarrow{\text{D}}$ biv. normal$((0,0), \Sigma^{**})$.

Example 5.13 (cont.) Recall the multinomial vector $(Y_1, \ldots, Y_k)$ described previously.

Now suppose we would like to estimate the *conditional* probability for category j, given the category is one of a proper subset $I \subset \{1, \ldots, k\}$ that includes j. This is $\frac{p_j}{\sum_{m \in I} p_m}$. The obvious estimator is $\frac{Y_j}{\sum_{m \in I} Y_m}$.

For simplicity (and because the problem can effectively be reduced to this case), suppose $j = 1$, $I = \{1, 2\}$ and $k = 3$. We start by noting that

$$\sqrt{n}(Y_1/n - p_1, Y_2/n - p_2) \xrightarrow{\text{D}} \text{biv. normal}((0, 0), \Sigma),$$

with

$$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1 p_2 \\ -p_1 p_2 & p_2(1-p_2) \end{pmatrix}.$$

(Note that $p_1 + p_2 < 1$, so Σ is invertible and the limit distribution has a joint bivariate normal pdf.)

Let $g(x_1, x_2) = \frac{x_1}{x_1 + x_2}$. Our objective, then, is to get an asymptotic distribution for $g(Y_1/n, Y_2/n) = \frac{Y_1/n}{Y_1/n + Y_2/n}$ as an estimator of $g(p_1, p_2) = \frac{p_1}{p_1 + p_2}$.

We will again use the delta method, but will just get one variance. The method states that we need

$$\frac{\partial g}{\partial x_1} = \frac{x_2}{(x_1+x_2)^2} \quad \text{and} \quad \frac{\partial g}{\partial x_2} = -\frac{x_1}{(x_1+x_2)^2}.$$

The variance we need is

$$\begin{aligned}
\sigma_*^2 &= \left[\left(\frac{\partial g}{\partial x_1}\right)^2 \sigma_{1,1} + 2\left(\frac{\partial g}{\partial x_1}\right)\left(\frac{\partial g}{\partial x_2}\right)\sigma_{1,2} + \left(\frac{\partial g}{\partial x_2}\right)^2 \sigma_{2,2}\right]_{x_1=p_1, x_2=p_2} \\
&= \frac{p_2^2 p_1(1-p_1)}{(p_1+p_2)^4} - \frac{2p_1^2 p_2^2}{(p_1+p_2)^4} + \frac{p_1^2 p_2(1-p_2)}{(p_1+p_2)^4} \\
&= \frac{p_1 p_2}{(p_1+p_2)^3}.
\end{aligned}$$

(Check: this is symmetric in the two variables, which is expected since a variance for $\frac{Y_1/n}{Y_1/n+Y_2/n}$ ought to equal that for $\frac{Y_2/n}{Y_1/n+Y_2/n} = 1 - \frac{Y_1/n}{Y_1/n+Y_2/n}$.)

Therefore,

$$\sqrt{n}\left(\frac{Y_1/n}{Y_1/n+Y_2/n} - \frac{p_1}{p_1+p_2}\right) \xrightarrow{\text{D}} \text{normal}\left(0, \frac{p_1 p_2}{(p_1+p_2)^3}\right).$$

This, then, can be used to develop statistical methods for inference about $\frac{p_1}{p_1+p_2}$.

Example 5.14 (cont.) Continue to assume that we have iid $X_i \sim \text{gamma}(\alpha, \beta)$. This time we will work with the fact that gamma(α, β) is a two-parameter exponential family. Specifically, the pdf is

$$f_X(x) = \frac{\mathrm{e}^{(\alpha-1)\log(x) - x/\beta}}{\beta^\alpha \Gamma(\alpha)}.$$

We already know that $\mathsf{E}(X) = \alpha\beta$. But here we will also derive $\mathsf{E}(\log(X))$ using the method of Thm. 3.18. (Recall also Ex. 5.11.) For this purpose we require the digamma function $\Psi(\alpha) = \frac{\partial \log(\Gamma(\alpha))}{\partial \alpha} = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$.

Taking derivatives with respect to each *parameter*,

$$0 = \frac{\partial}{\partial \alpha} \int_0^\infty f_X(x)\,\mathrm{d}x = \int_0^\infty \Big(\log(x) - \log(\beta) - \Psi(\alpha)\Big) f_X(x)\,\mathrm{d}x$$

so that $\mathsf{E}(\log(X)) = \log(\beta) + \Psi(\alpha)$, and

$$0 = \frac{\partial}{\partial \beta} \int_0^\infty f_X(x)\,\mathrm{d}x = \int_0^\infty \Big(\frac{-\alpha}{\beta} + \frac{x}{\beta^2}\Big) f_X(x)\,\mathrm{d}x$$

so that $\mathsf{E}(X) = \alpha\beta$.

Now, for each of the equations above, take another derivative with respect to each parameter. Hence,

$$0 = \frac{\partial^2}{\partial \alpha^2} \int_0^\infty f_X(x)\,\mathrm{d}x = \cdots = \mathsf{E}((\log(X) - \log(\beta) + \Psi(\alpha))^2) - \Psi'(\alpha),$$

$$0 = \frac{\partial^2}{\partial \beta^2} \int_0^\infty f_X(x)\,\mathrm{d}x = \cdots = \mathsf{E}\Big(\Big(\frac{(X-\alpha\beta)^2}{\beta^4}\Big) + \frac{\alpha}{\beta^2} - \frac{2X}{\beta^3}\Big),$$

and

$$\begin{aligned} 0 &= \frac{\partial^2}{\partial \alpha \partial \beta} \int_0^\infty f_X(x)\,\mathrm{d}x = \cdots \\ &= \mathsf{E}\Big(\frac{(\log(X) - \log(\beta) - \Psi(\alpha))(X - \alpha\beta)}{\beta^2}\Big) - \frac{1}{\beta}, \end{aligned}$$

The first equation gives $\mathsf{var}(\log(X)) = \Psi'(\alpha)$, the second gives $\mathsf{var}(X) = \alpha\beta^2$ and the third gives $\mathsf{cov}(\log(X), X) = \beta$.

We conclude, finally, that

$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n \log(X_i) - \mathsf{E}(\log(X)), \overline{X} - \mathsf{E}(X)\Big) \xrightarrow{\mathrm{D}} \mathsf{normal}((0,0), \Sigma),$$

where the elements of Σ are the variances and covariance above.

5.6 Empirical CDF and Quantiles

While means and moments get the most attention, estimating the distribution function itself and its quantiles are just as important.

DEFINITION 5.24 Let $X_1, \ldots, X_n$ be a simple random sample from a distribution F. The empirical cdf is the function

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i) = \text{proportion of data not more than } x.$$

- The empirical cdf is a *step function* that corresponds to a discrete distribution with jumps at the points $X_1, \ldots, X_n$. The size of each jump is the proportion of data that have a given value. See Fig. 5.6.
- If the data are continuous, the empirical cdf for such a sample will have jumps only of size $1/n$ since independent continuous rvs never take (exactly) the same value more than once.
- But "ties" can and do occur with discrete data (or rounded data) so in that case the jumps can be any value of the form j/n where $j = 1, \ldots, n$.

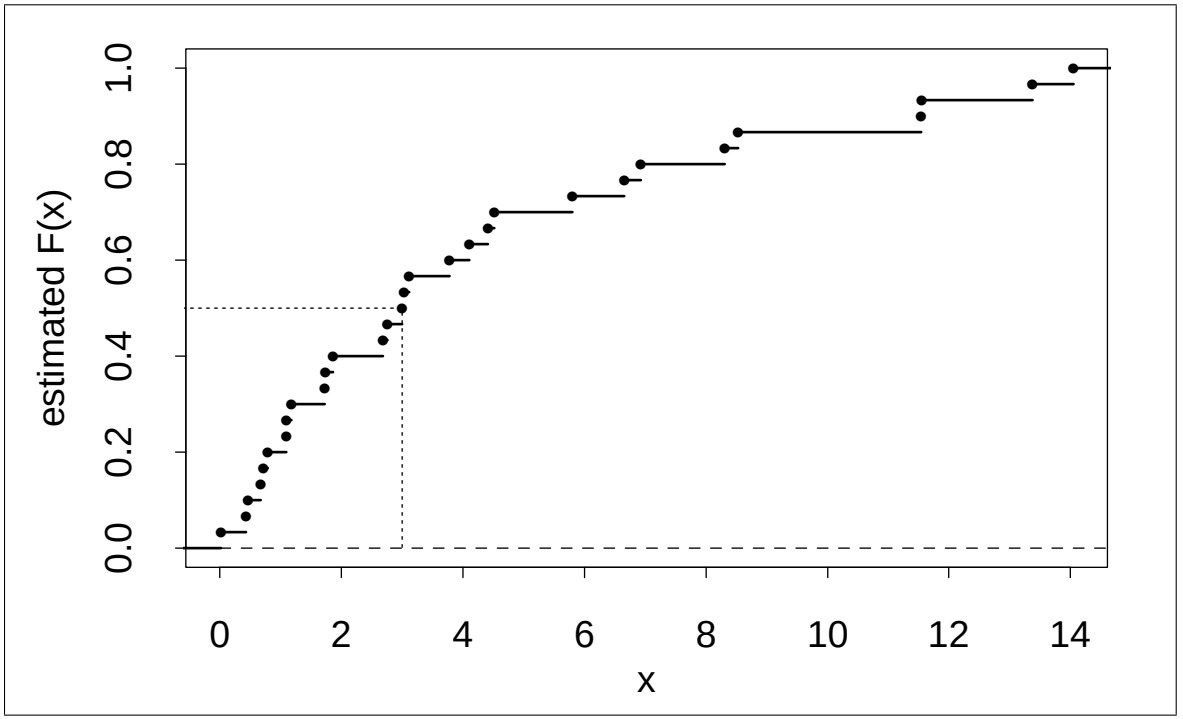

Figure 5.6 An empirical cdf for a sample of size 30, also identifying the median.

We mention, in passing, the following related problems.

- Estimating the pmf for discrete data is straightforward. One simply observes the proportion of data for each value of the random variable (which are the jump sizes of the empirical distribution). As proportions, the statistical properties are also easily derived.
- Estimating a pdf is somewhat harder. The simplest approach is by use of a histogram. But this is "blocky" looking, not smooth, and depends very much on the choice of intervals. However, there are now well-established techniques for "smoothing" that give rise to reasonable estimators of the pdf.

Since $\widehat{F}(x)$ is a *sample proportion* for each x, we immediately have the following results about its sampling distribution.

THEOREM 5.25 Assume as in Def. 5.24 and let $\widehat{F}$ be the empirical cdf.

i. For each real x, $n\widehat{F}(x)$ is a rv with binomial$(n, F(x))$ distribution.
In particular, $\mathrm{E}(\widehat{F}(x)) = F(x)$ and $\mathrm{var}(\widehat{F}(x)) = \frac{1}{n}F(x)(1-F(x))$.

ii. For each real x, $\widehat{F}(x) \to F(x)$ w.p. 1 and in m.s., as $n \to \infty$.

iii. For each real x, $\sqrt{n}\left(\frac{\widehat{F}(x)-F(x)}{\sqrt{F(x)(1-F(x))}}\right) \xrightarrow{\mathrm{D}}$ normal$(0,1)$ in distribution.

PROOF These all follow from the fact that $\widehat{F}(x)$ is a sample proportion.

i. $n\widehat{F}(x)$ is a sum of terms of the form $1_{(-\infty,x]}(X_i)$, each of which is a Bernoulli rv with success probability $\mathrm{P}(X_1 \leq x) = F(x)$. Thus, $n\widehat{F}(x)$ has binomial distribution.

ii. $\widehat{F}(x)$ is an average of iid random variables so the SLLN applies.

iii. See Ex. 5.8 for the CLT applied to sample proportions. □

Thm. 5.25 describes the sampling distribution of $\widehat{F}(x)$ for a single x. Although beyond the scope of this course, it is also of great interest to look at the joint distribution for multiple points, or even for the function as a whole.

- If $x_1 < x_2 < \cdots < x_k$ then the random vector

$$(n\widehat{F}(x_1), n(\widehat{F}(x_2) - \widehat{F}(x_1)), \ldots, n(1 - \widehat{F}(x_k)))$$

  has a *multinomial distribution*.

- For $x_1 < x_2$,

$$\text{cov}(\widehat{F}(x_1), \widehat{F}(x_2)) = \frac{F(x_1)(1 - F(x_2))}{n}.$$

- If $x_1 < x_2 < \cdots < x_k$ then $(\widehat{F}(x_1), \widehat{F}(x_2), \ldots, \widehat{F}(x_k))$ has *asymptotically multivariate normal distribution*, with means, variances and covariances as described above.

- $\max_{-\infty<x<\infty} |\widehat{F}(x) - F(x)| \to 0$ as $n \to \infty$, with probability 1. (That is, $\widehat{F}(x)$ converges *uniformly*.)

The sample quantile function is the inverse of the empirical distribution.

DEFINITION 5.26 Suppose $\widehat{F}$ is an empirical cdf. The sample p-th quantile is the p-th quantile of $\widehat{F}$, denoted $\widehat{x}_p$. In other words, it is the smallest value of x for which $\widehat{F}(x) \geq p$.

See Fig. 5.6 for an example of identifying the sample median, $\widehat{x}_{.50}$.

THEOREM 5.27 Assume as in Def. 5.24 and let $\widehat{F}$ be the empirical cdf. Assume also that F is a continuous distribution. Let $B(y; n, p)$ denote the binomial(n, p) cdf.

i. The cdf for $\widehat{x}_p$ is given by $\mathrm{P}(\widehat{x}_p \leq x) = 1 - B(\lceil np - 1 \rceil; n, F(x))$, where $\lceil y \rceil$ is the smallest integer greater than or equal to y.

ii. $\widehat{x}_p \to x_p$ w.p. 1 and in m.s.

iii. For large n, $\mathrm{P}(\widehat{x}_p \leq x) \doteq 1 - \Phi\left(\frac{\sqrt{n}(p - F(x))}{\sqrt{F(x)(1-F(x))}}\right)$.

PROOF

i. $\widehat{x}_p \leq x$ iff $p \leq \widehat{F}(x)$ and $Y = n\widehat{F}(x)$ has binomial$(n, F(x))$ distribution by Thm. 5.25.i, so $\mathsf{P}(\widehat{x}_p \leq x) = \mathsf{P}(n\widehat{F}(x) \geq np)$ can be obtained from the binomial distribution.

ii, iii. These follow from i. and the normal approximation to the binomial.

□

Applying the normal approximation from Thm. 5.27(iii) is not recommended unless the sample size is quite large. Using *the exact distribution* given in Thm. 5.27(i), which relies on binomial probabilities, is therefore recommended.

This theorem motivates the sign test, as in the following example.

Example 5.10 (cont.) Thirty ozone measurements of indeterminate distribution. Suppose the standard is that the *median* level be no more than 10 ppm, instead of the mean level, perhaps because the distribution appears to be skewed.

Suppose also, that only 12 of the data (40% of $n = 30$) are no more than 10 ppm (and thus $\widehat{x}_{.5} > 10$, suggesting that the standard is not met).

Let Y be the number of data no more than 10 ppm for a random sample. If in fact the standard is met so that $x_{.50} \leq 10$ then we would expect Y to be at least 15.

Specifically if $x_{.50} = 10$ (worst case) then $Y \sim$ binomial$(30, .5)$ and the chance we would get only 12 (as opposed to at least 15) is

$$\mathsf{P}(Y \leq 12) = 0.18080.$$

So the data are actually not all that unusual when the standard is met. (Compare, for example, to the chance of rolling one die and getting a 6.)

The normal approximation of Ex. 5.8 gives a less accurate value

$$\mathsf{P}(Y \leq 12) \doteq \Phi\left(\frac{\sqrt{n}(.40-.50)}{\sqrt{.50(.50)}}\right) = .137,$$

although there are methods to improve this approximation some.

5.7 Order Statistics and Extremes

The extreme statistics are the largest and smallest values in the data. Although they are not "central" and therefore do not represent the main part of the distribution, they are very important for understanding the tails of the distribution.

In particular, the maximum (the very largest datum) is studied in order to understand extreme phenomena such as rainfall, pollution levels, earthquakes, insurance claims, etc.

THEOREM 5.28 Suppose $X_1, \ldots, X_n$ is an iid sample from distribution $F(x)$. Then the cdf for $\max(X_1, \ldots, X_n)$ is $F^n(x)$.

Likewise, the cdf for $\min(X_1, \ldots, X_n)$ is $1 - (1 - F(x))^n$.

PROOF Observe that $\max(X_1, \ldots, X_n) \leq x$ iff each $X_i \leq x$. By independence,

$$\mathsf{P}(\max(X_1, \ldots, X_n) \leq x) = \prod_{i=1}^{n} \mathsf{P}(X_i \leq x) = F^n(x).$$

The case for the minimum is similar, using the fact that $\min(X_1, \ldots, X_n) > x$ iff each $X_i > x$. □

Example 5.1 (cont.) Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ exponential(β), which has cdf $F(x) = 1 - \mathrm{e}^{-x/\beta}$.

Then $\max(X_1, \ldots, X_n)$ has cdf $(1 - \mathrm{e}^{-x/\beta})^n$ and $\min(X_1, \ldots, X_n)$ has cdf $1 - \mathrm{e}^{-nx/\beta}$, which is another exponential distribution.

Example 5.15 Let $W_1, \ldots, W_n \overset{\text{iid}}{\sim}$ Weibull(γ, β) and $Y = \min(W_1, \ldots, W_n)$.

Then the cdf for Y is

$$F_Y(y) = 1 - (1 - F_W(y))^n = 1 - (\mathrm{e}^{-y^\gamma/\beta})^n = 1 - \mathrm{e}^{-ny^\gamma/\beta}, \quad y > 0,$$

which is the Weibull$(\gamma, \frac{\beta}{n})$ cdf.

Example 5.16 Let $T_1, \ldots, T_n$ be independent random variables with the Frechét(α, β) distribution $F_T(t) = \mathrm{e}^{-t^{-\alpha}/\beta}$ for $t > 0$.

Let $V = \max(T_1, \ldots, T_n)$. Then the cdf for V is

$$F_V(v) = (F_T(v))^n = \mathrm{e}^{-nv^{-\alpha}/\beta}, \quad v > 0,$$

which is the Frechét$(\alpha, \frac{\beta}{n})$ cdf.

Compare with the previous example, noting that if $T_i \sim$ Frechét(α, β) then $W_i = T_i^{-1} \sim$ Weibull(α, β), and $Y = V^{-1}$.

DEFINITION 5.29 The Order Statistics of an iid random sample $X_1, \ldots, X_n$ ($n > 1$) are the sample values placed *in order* from smallest to largest and denoted $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$.

- The order statistics are *not independent* (because they are in order) nor are they identically distributed.
- On the other hand, from a statistical point of view, they contain all *pertinent knowledge* about the distribution that the original data were sampled from. (This point assumes an iid random sample.)

If the distribution is discrete then there is positive probability of exactly equal values (ties). This makes things complicated so we will focus only on the continuous case.

Because there are $n!$ possible orderings of the data, it is a simple matter to see that the *joint pdf* for $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ must be

$$n! f(x_1) f(x_2) \cdots f(x_n), \quad x_1 < x_2 < \cdots < x_n.$$

Now we determine the marginal distributions of order statistics.

THEOREM 5.30 Suppose $X_1, \ldots, X_n$ is a random sample from continuous distribution $F(x)$ with pdf $f(x)$.

Then the cdf for $X_{(i)}$ is

$$F_{X_{(i)}}(x) = \sum_{j=i}^{n} \binom{n}{j} (F(x))^j (1 - F(x))^{n-j},$$

and the pdf for $X_{(i)}$ is

$$f_{X_{(i)}}(x) = n \binom{n-1}{i-1} f(x) (F(x))^{i-1} (1 - F(x))^{n-i}.$$

Heuristically, one can explain the pdf as follows: from the original $X_1, \ldots, X_n$, there are n variables that could be $X_{(i)} = x$ and $\binom{n-1}{i-1}$ ways to choose the $i-1$ variables that are less than x (and, consequently, the $n-i$ variables that are greater than x).

PROOF Fix x and note that data values are not equal to x with probability 1, since the distribution is continuous. Let Y be the number of data values less than x. Then Y has binomial$(n, F(x))$ distribution.

Moreover, $X_{(i)} \leq x$ iff $Y \geq i$. Hence,

$$F_{X_{(i)}}(x) = \mathsf{P}(X_{(i)} \leq x) = \mathsf{P}(Y \geq i) = \sum_{j=i}^{n} \binom{n}{j} (F(x))^j (1 - F(x))^{n-j}.$$

Taking a derivative with respect to x leads to a sum where like-terms cancel, leaving only the term shown for the pdf in the theorem statement above. □

- Note that Thm. 5.27.i and Thm. 5.30 are closely related (take $i = \lceil np \rceil$).
- Also, Thm. 5.28 describes the special cases of Thm. 5.30 with $i = n$ and $i = 1$, respectively.

Joint distributions of a subset of the order statistics can be obtained in similar fashion by considering the number of X_i's found in different intervals. Here, we will just focus on the joint distribution of the two extremes.

THEOREM 5.31 Assume as before.

i. The joint pdf for $(X_{(1)}, X_{(n)})$ is

$$f_{X_{(1)},X_{(n)}}(x_1, x_n) = n(n-1)f(x_1)f(x_n)(F(x_n) - F(x_1))^{n-2}.$$

ii. Let $R = X_{(n)} - X_{(1)}$ be the range of the sample and $M = \frac{X_{(1)}+X_{(n)}}{2}$ be the mid-range. Then

$$f_{R,M}(r, s) = n(n-1)f(s - r/2)f(s + r/2)(F(s + r/2) - F(s - r/2))^{n-2}.$$

Using the same heuristic described for Thm. 5.30, one can remember this joint pdf by noting that there are $n(n-1)$ choices for which X_i's are the min and max, and the probability that the rest are between x_1 and x_n is $(F(x_n) - F(x_1))^{n-2}$.

PROOF i. Observe that

$$\mathsf{P}(x_1 < X_{(1)} < X_{(n)} \leq x_n) = (F(x_n) - F(x_1))^n.$$

Taking derivatives of this with respect to each variable gives the *negative* of the joint pdf. (Why?)

ii. Observe that $X_{(1)} = M - \frac{1}{2}R$ and $X_{(n)} = M + \frac{1}{2}R$, with Jacobian determinant $\frac{\mathrm{d}x_1\,\mathrm{d}x_n}{\mathrm{d}r\,\mathrm{d}s} = 1$.

The result then follows from the methods for (linear) transformations of random variables (Thm. 4.18). □

Example 5.17 Let $V_1, \ldots, V_n \overset{\text{iid}}{\sim} \text{uniform}(a, b)$. Note that $V_{(i)} = a + (b-a)U_{(i)}$, where $U_{(1)}, \ldots, U_{(n)}$ are the order statistics for a uniform(0,1) random sample.

So we can easily derive results about the general case from the special uniform(0,1) case.

For that, we have $f(u) = 1$ and $F(u) = u$ for $0 \leq u \leq 1$. Thus,

$$f_{U_{(i)}}(u) = n\binom{n-1}{i-1}u^{i-1}(1-u)^{n-i},$$

which is the beta$(i, n-i+1)$ pdf.

Using what we know about the beta distribution,

$$\mathsf{E}(V_{(i)}) = a + (b-a)\mathsf{E}(U_{(i)}) = a + \frac{(b-a)i}{n+1},$$

and

$$\text{var}(V_{(i)}) = (b-a)^2\,\text{var}(U_{(i)}) = \frac{(b-a)^2 i(n-i+1)}{(n+1)^2(n+2)}.$$

In particular, $\mathsf{E}(V_{(1)}) = a + \frac{b-a}{n+1}$ and $\mathsf{E}(V_{(n)}) = b - \frac{b-a}{n+1}$ so that the expected range of the sample is $\mathsf{E}(V_{(n)} - V_{(1)}) = \frac{(n-1)(b-a)}{n+1}$, just shy of the data range $b-a$.

Let $R = U_{(n)} - U_{(1)}$ and $M = \frac{U_{(n)}+U_{(1)}}{2}$ be the range and mid-range, respectively, for the uniform(0,1) sample.

Keeping in mind that $0 < U_{(1)} < U_{(n)} < 1$ and therefore $\frac{R}{2} < M < 1 - \frac{R}{2}$, Thm. 5.31 says the joint pdf for (R, M) is

$$f_{R,M}(r, s) = n(n-1)r^{n-2}1_{[r/2,1-r/2]}(s)1_{[0,1]}(r).$$

It follows easily that

$$f_R(r) = \int_{r/2}^{1-r/2} n(n-1)r^{n-2}\,\mathrm{d}s = n(n-1)r^{n-2}(1-r), \quad 0 \le r \le 1,$$

which is the beta$(n-1, 2)$ pdf.

Since the restrictions imply $r \le 2\min(s, 1-s)$,

$$f_M(s) = \int_0^{2\min(s,1-s)} n(n-1)r^{n-2}\,\mathrm{d}r = n(2\min(s, 1-s))^{n-1}, \quad 0 \le s \le 1.$$

This last example suggests the following, which is an alternative way to characterize the distribution of an order statistic.

COROLLARY 5.32 Assume as in Thm. 5.31. Then $F(X_{(i)}) \sim \text{beta}(i, n-i+1)$.

PROOF Recall Thm. 2.35 and Thm. 2.37. Combined, these say that for continuous $F(x)$, $X \sim F \iff F(X) \sim \text{uniform}(0,1)$.

From this fact we can deduce that $F(X_{(i)}) = U_{(i)}$, where $U_{(1)}, \ldots, U_{(n)}$ are the order statistics for a uniform(0,1) random sample. The conclusion then follows from Ex. 5.17. □

Now we turn to *asymptotics* for the sample maximum. Note that similar results will hold for the sample minimum (by replacing X_i with either $-X_i$ or $1/X_i$, for example).

It turns out that there are *three separate cases* – which we will simplify slightly for the sake of illustration. For the rest of this section, we assume $X_1, X_2, \ldots$ is an iid random sequence from distribution $F(x)$, and we let $M_n = \max(X_1, \ldots, X_n)$ for each n.

Our objective is to get an asymptotic distribution for M_n, when n is large.

THEOREM 5.33 If $F(x)$ satisfies $\lim_{x\to\infty} x^\alpha(1-F(x)) = c > 0$, with $\alpha > 0$, then

$$\lim_{n\to\infty} \mathsf{P}(M_n \le n^{1/\alpha}x) = \mathrm{e}^{-cx^{-\alpha}},$$

which is the Frechét$(\alpha, 1/c)$ cdf (see Ex. 5.16).

PROOF First, recall that $\log(1-y) \sim -y$ as $y \to 0$. Using Thm. 5.28 and taking logarithms,

$$\log(\mathsf{P}(M_n \le n^{1/\alpha}x)) = n\log(F(n^{1/\alpha}x)) \sim -n(1-F(n^{1/\alpha}x)) \sim -cx^{-\alpha}.$$

Reverting back by taking exponentials gives the result. □

The distributions that fit the case of Thm. 5.33 are said to have heavy-tails because $1-F(x)$ decreases more slowly than exponentially.

Typical examples include the Pareto distribution: $F(x) = 1 - cx^{-\alpha}$ for $x > c^{1/\alpha}$, the Lomax distribution: $F(x) = 1-(1+x/\beta)^{-\alpha}$, and the Frechét$(\alpha,\beta)$ distribution itself.

At the other end of the spectrum are distributions for random variables that have a finite upper bound b. This includes uniform and beta distributions. The second result is concerned with those types.

THEOREM 5.34 If $F(x)$ satisfies $\lim_{x\to 0} x^{-\gamma}(1-F(b-x)) = c > 0$, with $\gamma > 0$, then

$$\lim_{n\to\infty} \mathsf{P}(b - M_n \le n^{-1/\gamma} x) = 1 - \mathrm{e}^{-cx^\gamma}.$$

which is the Weibull$(\gamma, 1/c)$ distribution.

PROOF (Exercise. Either follow logic similar to the proof of Thm. 5.33, with appropriate adjustments, or consider that the sequence $T_i = \frac{1}{b-X_i}$ satisfies the conditions for Thm. 5.33, and then use the remark following Ex. 5.16.) □

Example 5.17 (cont.) Let $V_1, V_2, \ldots \overset{\text{iid}}{\sim}$ uniform(a, b). Here, $1 - F(b-x) = \frac{x}{b-a}$ for $a \le x \le b$ and we see that we can take $\gamma = 1$ and $c = \frac{1}{b-a}$. Therefore, with $M_n = \max(V_1, \ldots, V_n)$, $\mathsf{P}(b - M_n \le x/n) \to 1 - \mathrm{e}^{-x/(b-a)}$, which is an exponential distribution. We can approximate

$$\mathsf{P}(M_n \le y) \approx \mathrm{e}^{-n(b-y)/(b-a)}, \quad \text{for } y < b.$$

The third case is intermediate to the previous two cases and, while it includes common distributions such as normal, gamma and Weibull, it is *much trickier*. These are the cases where there is no upper bound but the tails decrease faster than a power function.

Thus we will just look at a couple of examples.

Example 5.15 (cont.) Suppose $W_1, W_2, \ldots \overset{\text{iid}}{\sim} F = \text{Weibull}(\gamma, 1)$ and let $M_n = \max(W_1, \ldots, W_n)$.

The objective is to find sequences of constants, a_n and b_n, such that $\frac{M_n - b_n}{a_n}$ converges in distribution. That is,

$$\text{P}(M_n \leq a_n x + b_n) = (F(a_n x + b_n))^n \to G(x),$$

for some cdf $G(x)$.

Note that $a_n x + b_n \to \infty$ is required for such convergence since M_n must also be increasing to infinity.

Then, taking logarithms as the previous cases,

$$\log(\text{P}(M_n \leq a_n x + b_n)) \sim -n(1 - F(a_n x + b_n)) = -\text{e}^{-(a_n x + b_n)^\gamma + \log(n)}.$$

We want what is in the exponent to have a limit that depends on x.

Let $c_n = \frac{b_n}{a_n} = (\log(n))^{1/\gamma}$ so that

$$(a_n x + b_n)^\gamma - \log(n) = a_n^\gamma \int_{c_n}^{c_n+x} \gamma u^{\gamma-1}\,\mathrm{d}u \sim \gamma a_n^\gamma c_n^{\gamma-1} x.$$

Putting $a_n^\gamma c_n^{\gamma-1} = 1$ gives $a_n = (\log(n))^{(1-\gamma)/\gamma^2}$ and $b_n = (\log(n))^{1/\gamma^2}$.

Therefore,

$$\mathsf{P}(M_n \leq a_n x + b_n) \to G(x;\gamma) = \mathrm{e}^{-\mathrm{e}^{-\gamma x}}.$$

The limit is known as a Gumbel distribution.

In summary, we have the approximation $\mathsf{P}(M_n \leq y) \approx G(\frac{y-b_n}{a_n};\gamma)$.

Example 5.14 (cont.) Suppose $T_1, T_2, \ldots \overset{\text{iid}}{\sim} F = \text{gamma}(\alpha, \beta)$ and let

$$M_n = \max(T_1, \ldots, T_n).$$

Once again, we want a_n and b_n such that

$$-n(1 - F(a_n x + b_n)) \to \log(G(x)).$$

We start by observing that $1 - F(x) \sim \frac{(x/\beta)^{\alpha-1}\mathrm{e}^{-x/\beta}}{\Gamma(\alpha)}$ as $x \to \infty$.

Then,

$$-n(1 - F(a_n x + b_n)) \sim \frac{n((a_n x + b_n)/\beta)^{\alpha-1}\mathrm{e}^{-(a_n x + b_n)/\beta}}{\Gamma(\alpha)}.$$

Here is where it takes a little guesswork. As we want the limit to depend on x and the exponential part dominates, we try $a_n = \beta$ for all n and then suppose $n(b_n/\beta)^{\alpha-1}\mathrm{e}^{-b_n/\beta} \to 1$.

That is, $b_n \sim \beta(\log(n) - (\alpha - 1)\log\log(n))$. With those choices we get

$$\mathsf{P}(M_n \leq a_n x + b_n) \to G(x) = \mathrm{e}^{-\mathrm{e}^{-x}}.$$

As in the previous example, the limit is a Gumbel distribution.

More generally, assuming $F(x)$ is strictly increasing and continuous, the limit will be Gumbel if one can choose

$$b_n \sim F^{-1}(1-1/n) \quad \text{and} \quad a_n \sim F^{-1}(1-\mathrm{e}^{-1}/n) - F^{-1}(1-1/n),$$

with the proviso that $\frac{a_{2n}-a_n}{b_n}$ converges to some constant.

5.8 Counting Rare Events and Compound Poisson RVs

We return to the problem of sums of Bernoulli trials except we now we consider so-called *rare events* with probabilities that decrease as the sample size n grows larger. Like extremes, this is a topic that has generated more interest in recent decades. It has applications for traffic, networks, ecology, geology, and genetics, among other areas of research.

Rare events change the model in such a way that the CLT is no longer applicable. Instead, we get distributional convergence of a sum without any normalization.

Additionally, we define a new sample for each n. Specifically, let $(X_{n,1}, \ldots, X_{n,n})$ denote a sample of n *independent* Bernoulli trials with *possibly different* success probabilities $p_{n,i} = \mathsf{E}(X_{n,i})$. Also set $S_n = \sum_{i=1}^{n} X_{n,i}$ and

$$\lambda_n = \mathsf{E}(S_n) = \sum_{i=1}^{n} p_{n,i}.$$

Events (successes) are "rare" if the $p_{n,i}$'s are uniformly very small. In this case, the expected number of successes λ_n cannot grow but does stay positive.

Before we continue, we recall a special case that was investigated in Ex. 2.18. If the success probabilities within a sample are the same (i.e., $p_{n,i} = p_n$ for all i) then $S_n \sim$ binomial(n, p_n). Suppose, in addition, that $\lambda_n = np_n \to \lambda > 0$, as $n \to \infty$. For this case, we showed

$$S_n \xrightarrow{\text{D}} \text{Poisson}(\lambda).$$

Now we generalize by letting the probabilities vary while still being very small.

THEOREM 5.35 (Law of Rare Events) Assume $(X_{n,i}, \ldots, X_{n,n})$ and S_n are as above.

If $\lambda_n \to \lambda \in (0, \infty)$ and $\delta_n = \max_{i \le n} p_{n,i} \to 0$, as $n \to \infty$, then

$$S_n \xrightarrow{\text{D}} \text{Poisson}(\lambda).$$

Note: there is *no law of large numbers* here. This is strictly about convergence in distribution. (Indeed, there is nothing assumed about how each S_n is related to any other S_m.)

PROOF Just as in other proofs of this nature, we show the desired distributional convergence *by showing convergence of mgfs* (Thm. 2.32). To be precise, we want to show

$$\log(\mathsf{E}(\mathrm{e}^{tS_n})) \to \lambda(\mathrm{e}^t - 1),$$

which is the logarithm of the Poisson(λ) mgf.

We will use the fact that $|\log(1+x) - x| \leq x^2$ for $|x| \leq \frac{1}{2}$. Observe first that

$$\log(\mathsf{E}(\mathrm{e}^{tS_n})) = \log\Big(\prod_{i=1}^{n} \mathsf{E}(\mathrm{e}^{tX_{n,i}})\Big) = \sum_{i=1}^{n} \log(1 + p_{n,i}(\mathrm{e}^t - 1)).$$

Then,

$$\begin{aligned}
|\log(\mathsf{E}(\mathrm{e}^{tS_n})) - \lambda(\mathrm{e}^t - 1)| &= \Big|\sum_{i=1}^{n} \log(1 + p_{n,i}(\mathrm{e}^t - 1)) - \lambda(\mathrm{e}^t - 1)\Big| \\
&\leq \sum_{i=1}^{n} |\log(1 + p_{n,i}(\mathrm{e}^t - 1)) - p_{n,i}(\mathrm{e}^t - 1)| + |(\lambda_n - \lambda)(\mathrm{e}^t - 1)| \\
&\leq \sum_{i=1}^{n} (p_{n,i}(\mathrm{e}^t - 1))^2 + |(\lambda_n - \lambda)(\mathrm{e}^t - 1)| \\
&\leq \delta_n \lambda_n (\mathrm{e}^t - 1)^2 + |(\lambda_n - \lambda)(\mathrm{e}^t - 1)| \to 0.
\end{aligned}$$

This shows that $\mathsf{E}(\mathrm{e}^{tS_n})$ converges to the Poisson(λ) mgf, as required. □

Example 5.18 A traffic engineer has historical data that gives the probability of an accident each hour of the day for a busy intersection. The probabilities vary quite a bit over the different hours of the day. They are all small but add up to a not too small a value of 0.15. Over a 30-day period this would be an expected total of 4.5 accidents.

The law of rare events therefore suggests that the actual number of accidents in a 30-day period can be modeled with a Poisson(4.5) distribution.

Two other busy intersections can likewise be modeled with Poisson distributions having means 2.6 and 1.8, respectively, per 30 days. As accidents at the three intersections ought to occur independently, the total number for all three in a given 30-day period will have a Poisson(8.9) distribution.

This idea can be extended to the notion of a Poisson point process where the times of accidents are *random points in time* but the total number of points over a given interval has a Poisson distribution. An additional assumption, reasonable here, is that disjoint intervals of time are independent with regard to the number of accidents in each.

Example 5.19 A network has n nodes, where n is very large. Suppose $W_1, \ldots, W_n$ are iid random, positive *weights* with finite mean $\mu = \mathsf{E}(W)$. Let $\overline{W}$ be the average weight which, of course, converges to μ by the SLLN. One other consequence of the SLLN is that

$$\frac{1}{n} \max_{j \le n} W_j \to 0, \text{ as } n \to \infty.$$

Nodes i and j, $i \neq j$, form an *edge* (connection) with probability $\frac{W_i W_j}{n\overline{W}}$, conditional on the weights. Let $E_{i,j}$ be the indicator of such an edge. Define $D_i = \sum_{j \neq i} E_{i,j}$ to be the *degree* (number of edges) that node i has. Observe that

$$\mathsf{E}(D_i | W_1, \ldots, W_n) = \sum_{j \neq i} \frac{W_i W_j}{n\overline{W}} = W_i \frac{n\overline{W} - W_i}{n\overline{W}} \approx W_i.$$

By Thm. 5.35, we determine that D_i has approximately Poisson(W_i) distribution, *conditional on the weights*. Indeed, conditionally,

$$D_i \xrightarrow{\mathrm{D}} \text{Poisson}(W_i), \text{ as } n \to \infty.$$

Suppose D has the *unconditional* limit distribution. Then its mgf is

$$M_D(t) = \mathsf{E}(\mathsf{E}(\mathrm{e}^{tD} | W)) = \mathsf{E}(\mathrm{e}^{W(\mathrm{e}^t - 1)}) = M_W(\mathrm{e}^t - 1).$$

Now we return to the model with independent Bernoulli trials $(X_{n,1}, \ldots, X_{n,n})$ with respective success probabilities $p_{n,i} = \mathsf{E}(X_{n,i})$, $S_n = \sum_{i=1}^n X_{n,i}$ and $\lambda_n = \mathsf{E}(S_n) = \sum_{i=1}^n p_{n.i}$.

Suppose, in addition that there is an iid sequence $Y_1, Y_2, \ldots$ (called *marks*), independent of $\{X_{n,i}\}$. For fixed n, the only marks Y_i that are observed/actualized are those with $X_{n,i} = 1$. Define $T_n = \sum_{i=1}^n X_{n.i} Y_i$, which is the total of the observed marks.

COROLLARY 5.36 Under the assumptions above, T_n converges in distribution to the distribution with mgf $\mathrm{e}^{\lambda(M_Y(t)-1)}$.

PROOF Observe that

$$\mathsf{E}(\mathrm{e}^{tX_{n,i}Y_i}) = p_{n,i}\mathsf{E}(\mathrm{e}^{tY_i}) + (1-p_{n,i})\mathsf{E}(\mathrm{e}^0) = 1 + p_{n,i}(M_Y(t)-1).$$

Following the proof of Thm. 5.35 almost exactly, we get

$$\mathsf{E}(\mathrm{e}^{tT_n}) \to \mathrm{e}^{\lambda(M_Y(t)-1)}, \text{ as } n \to \infty.$$

This proves the result. □

On the face of it, T_n is asymptotically the sum of a Poisson(λ) number of iid terms. Such a sum has a compound Poisson distribution. To confirm, suppose $N \sim \text{Poisson}(\lambda)$, independent of $Y_1, Y_2, \ldots$, and let $T = \sum_{i \le N} Y_i$ (with $T = 0$ if $N = 0$). Then by partitioning on the value of N,

$$\mathsf{E}(\mathrm{e}^{tT}) = \sum_{n=0}^{\infty} \mathsf{E}(\mathrm{e}^{t(Y_1+\cdots+Y_n)})\mathsf{P}(N=n) = \sum_{n=0}^{\infty} (\mathsf{E}(\mathrm{e}^{tY}))^n \frac{\lambda^n \mathrm{e}^{-\lambda}}{n!} = \cdots = \mathrm{e}^{\lambda(M_Y(t)-1)}.$$

So, $T_n \xrightarrow{\mathrm{D}} T$.

Some values that are evident, either from the definition of T or from its mgf, are

$$\mathsf{E}(T) = \lambda \mathsf{E}(Y) \quad \text{and} \quad \mathsf{var}(T) = \lambda \mathsf{E}(Y^2)$$

(exercise).

Compound Poisson random variables have many applications. They are harder to analyze statistically if observed on their own. However, it is not unusual to observe a random sample $(Y_1, \ldots, Y_N)$ where the sample size N is itself random (with, for example, a Poisson distribution). Such a situation is more amenable to inference.

Example 5.20 The number of automobile accidents per month in a particular region, insured by American Insurance Company, can be modeled with a Poisson($\lambda = 35$) distribution. Collision repair claims, after deductible, have an exponential($\beta = \$850$) distribution.

The total T of monthly claims has the resulting compound Poisson distribution. In this case, we can find the distribution exactly, if not simply. Let N be the number of claims.

First, note that $\mathsf{P}(T = 0) = \mathsf{P}(N = 0) = \mathrm{e}^{-\lambda}$. So the distribution is a *mixture*, with one jump at $t = 0$. In this case, $\mathsf{P}(T = 0) = \mathrm{e}^{-35} = 6.305 \times 10^{-16}$, so the discrete part is not of much concern.

For the case $N > 0$, we note that the sum of n iid exponential rvs has gamma distribution and hence we get a "pdf" (that integrates to $1 - \mathrm{e}^{-\lambda}$)

$$f_T(t) = \sum_{n=1}^{\infty} \frac{\lambda^n \mathrm{e}^{-\lambda}}{n!} \frac{t^{n-1} \mathrm{e}^{-t/\beta}}{n! \beta^n} = \sum_{n=1}^{\infty} \frac{(\lambda t/\beta)^n \mathrm{e}^{-(t/\beta + \lambda)}}{(n!)^2}.$$

The cdf is thus

$$F_T(t) = \mathrm{e}^{-\lambda} + \int_0^t f_T(x)\,\mathrm{d}x, \quad \text{for } t \geq 0.$$

Example 5.21 Suppose rare events occur at points in the interval $[0,1]$ in such a way that the probability of at least one occurrence in the interval $[\frac{(i-1)}{n}, \frac{i}{n})$ (with indicator rv $X_{n,i}$) is $p_{n,i} \approx \frac{\lambda}{n}$, and such event is independent of what happens on other intervals.

Assume also that the probability of more than one occurrence in the small interval is negligible (that is, $o(1/n)$).

The total number S_n of occurrences in $[0,1]$ is asymptotically Poisson(λ), by Thm. 5.35.

Next, we assume also that each occurrence has a random categorical (or discrete) mark Y_i, which are iid with probabilities $q_1, \ldots, q_k$, $q_1 + \cdots + q_k = 1$. Let $T_{n,j}$ be the number of events with mark value j.

Observe that $T_{n,j} = \sum_{i=1}^{n} X_{n,i} 1_{Y_i=j}$ is the sum of independent Bernoulli($p_{n,i}q_j$) random variables. Thus, $T_{n,j} \xrightarrow{\text{D}}$ Poisson(λq_j).

On the other hand, the vector $(T_{n,1}, \ldots, T_{n,k})$ has a multinomial distribution for each n. What is the limiting *joint* distribution?

There several ways to solve this. One is to find the limit of the *joint mgf* $\mathsf{E}\left(\mathrm{e}^{t_1 T_{n,1} + \cdots + t_k T_{n,k}}\right)$.

Another is to observe, by combining several categories into one, that for any $I \subset \{1, \ldots, k\}$, $\sum_{j \in I} T_{n,j} \xrightarrow{\mathrm{D}} \mathsf{Poisson}(\lambda \sum_{j \in I} q_j)$.

Either method verifies that the limit distribution for $(T_{n,1}, \ldots, T_{n,k})$ must be that of *independent* Poisson rvs. Therefore, $T_{n,1}, \ldots, T_{n,k}$ are asymptotically independent.

