2. Introduction to statistics: first examples

2.1. **Introduction.** The basic problem of statistics is to draw conclusions about unknown distributions of random variables from observed values. These conclusions inevitably will have to be (you guessed it) statistical in nature: if you flip heads 100 times in a row, then perhaps both sides of this coin are heads, or it could be a perfectly fair coin and you just had a lucky run of heads, and obviously you can't tell for sure which explanation is right. In mathematical statistics, we try to deal with this uncertainty in a controlled and quantitative way, if we can.

For a typical situation, consider a random variable X with a distribution that depends on an unknown parameter θ. We can describe it by giving the probabilities $p_n(\theta) = P(X = x_n)$ (discrete case) or the density $f(x, \theta)$ (continuous case). Suppose we have iid random variables $X_1, \ldots, X_n$ with this distribution; in statistics, this is also called a *random sample.* The real life situation we have in mind here is that of a random experiment that we can repeat as many times as we wish, and the individual experiments are independent of each other.

Definition 2.1. Let $X_1, \ldots, X_n$ be a random sample. A *statistic* is a function $T = T(X_1, \ldots, X_n)$.

In other words, a statistic is a random variable itself, but one of a special form: it is a function of the random sample. The definition is completely general, but what we have in mind here is a more concrete situation where T serves as a guess on the unknown parameter θ: if we observe the values $x_1, \ldots, x_n$ of the random sample, then we guess $\theta = T(x_1, \ldots, x_n)$, and we are really interested in statistics T for which this guess is reasonable, in various senses that we will make precise later. If T is indeed used in this way to guess a parameter, we also call T an *estimator.*

Let's look at an example: Suppose $P(X_1 = 1) = \theta$, $P(X_1 = 0) = 1 - \theta$, with $0 \leq \theta \leq 1$. You can think of a biased coin with an unkown probability θ of coming up heads. You now observe a random sample of size n, or, in plain English, you toss this coin n times and observe the results $x_j = 0, 1$ for $j = 1, \ldots, n$. What would now be a good statistic $T = T(X_1, \ldots, X_n)$ if you want to take a guess on the value of θ? Obviously, there is only one sane way to go about it: we guess that

$$T = \frac{1}{n}(X_1 + \ldots + X_n). \tag{2.1}$$

To make the theory more mathematical, we will, over the course of the semester, compile various desirable properties that we would like our statistics to have. For example:

Definition 2.2. A statistic T is called *unbiased* if $ET = \theta$ for all θ.

So while an unbiased estimator may of course return incorrect values in any given case, things at least come out right on average, no matter what the unknown parameter is actually equal to. Note also that the expectation ET is taken with respect to the unknown, θ-dependent distribution, so we could have written $E_\theta T$ to emphasize this.

The estimator T from (2.1) is unbiased:

$$ET = \frac{1}{n}(EX_1 + \ldots + EX_n) = EX_1 = \theta$$

T has many other desirable properties, and we will return to this example many times.

While it is nice to have unbiased estimators, this criterion cannot be used uncritically. Consider the following example: you have an urn with an unknown number $N = \theta$ of balls in it, and these are numbered $1, 2, \ldots, N$. (I'll write N instead of θ for the unknown parameter because it feels more natural in this example.) You now draw balls, with replacement, n times from this urn, and you would like an estimator T for N, based on this random sample (please don't spoil the fun by asking why I don't draw *without* replacement until the urn is empty).

Let's make this more formal. Let X_j be the number of the jth ball drawn. Then $X_1, \ldots, X_n$ are iid (= a random sample), with common distribution $P(X_1 = m) = 1/N$, $m = 1, 2, \ldots, N$. If I draw sufficiently many times, I would normally hope to get the ball with the maximal label N at some point, so why don't we try

$$T = \max(X_1, \ldots, X_n)$$

as our statistic. Say I drew $1, 1, 2, 1, 1, 1, 3, 3, 2, 3, 1, 1, 1, 2$: then I'll guess $N = T(\ldots) = 3$.

Clearly, T is biased: we cannot get values $T > N$, but we do get values $T < N$. More precisely, $P(T < N) > 0$, so

$$\begin{aligned} ET &= \sum_{m=1}^{N} mP(T = m) \leq (N-1)P(T < N) + NP(T = N) \\ &= N - P(T < N) < N. \end{aligned}$$

However, T has redeeming qualities. For example, if our random sample is large enough, we guess correctly most of the time if we use T:

$$P(T = N) = 1 - P(T < N) = 1 - \left(\frac{N-1}{N}\right)^n \to 1 \qquad (n \to \infty)$$

This also confirms that while T does show a bias towards values that are too small, we have $ET \to N$ as $n \to \infty$, so this bias becomes very small for large samples.

We can also build an unbiased estimator. To do this, let's take a look at the *sample mean*

$$\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n).$$

We compute

$$E\overline{X} = EX_1 = \sum_{m=1}^{N} m\frac{1}{N} = \frac{N(N+1)}{2}\,\frac{1}{N} = \frac{N+1}{2}.$$

So $\overline{X}$ isn't unbiased either, but now we can pass to $U = 2\overline{X} - 1$, and this new estimator is unbiased since $EU = 2E\overline{X} - 1 = N$.

Of course, U does not feel right at all as an estimator for N; for example, repeatedly drawing balls with small numbers will make U small, but does not provide evidence for a small N. Despite being biased, T should clearly be preferred. (Later we will see that T has other highly desirable properties.)

Exercise 2.1. (a) Find the distribution of T, that is, find the probabilities $P(T = m)$, $m = 1, 2, \ldots, N$. *Suggestion:* Start out by working out $P(T \le m)$.
(b) Show that

$$ET = \sum_{m=1}^{N} m\left[\left(\frac{m}{N}\right)^n - \left(\frac{m-1}{N}\right)^n\right].$$

(c) Let's call this expression $\mu = \mu(N, n)$. Why don't we similarly "fix" T by passing to $V = NT/\mu$, so that $EV = N$ and V is unbiased? What is wrong with this?

Exercise 2.2. Draw the N balls *without* replacement, and define the random variable $Y_j = 1, \ldots, N$ as the jth ball drawn. Argue that the Y_j still have the same distribution $P(Y_j = m) = 1/N$, $m = 1, \ldots, N$, but they are not independent.

Exercise 2.3. Suppose you have a coin that is either fair ($p = \theta = 1/2$), or you have heads on both sides of the coin ($\theta = 1$). Then the following statistic T looks like a reasonable estimator for θ:

$$T = \begin{cases} 1 & X_1 = \ldots = X_n = 1 \\ 1/2 & \text{otherwise} \end{cases}$$

(a) Compute ET for both $\theta = 1/2$ and $\theta = 1$ and conclude that T is not unbiased.
(b) Now suppose you can also have tails on both sides of the coin ($\theta = 0$). Define an analogous estimator and show that this estimator is unbiased.
(c) Come to think of it, in general, if θ takes just two distinct values, do you really want to insist on unbiased estimators?

Let me conclude this section with a few somewhat sobering remarks. Let's go back to the coin flip example. To keep things nice and simple, let's in fact assume that the unknown probability θ takes only finitely many values $\theta = \theta_1, \ldots, \theta_N$. As before, we want to guess θ, based on the observation of a random sample. It is natural to now define $T(X_1, \ldots, X_n)$ as the θ_j that is closest to $\overline{X}$ (and take the larger value, say, if there are two such values).

So far, our analysis has focused on scenarios of the type: Suppose that θ has a certain value. What can then be said about the behavior of T? (Pretty much the only quantity we've looked at so far is ET, but we'll analyze many other questions of this type as we proceed.)

In a real life application, θ is exactly what we are ignorant about. So the question we really want to ask is (subtly, perhaps, but crucially) different from what we actually did. We want to know: "I just guessed that $\theta = \theta_j$, using the statistic T. I am aware that there is a possibility of error, but what is the probability that my guess was correct?" Or, in more mathematical terms, what is the conditional probability

$$P(\theta = \theta_j | T = \theta_j) \tag{2.2}$$

equal to?

Unfortunately, there isn't much we can do here. We *can* work out the conditional probabilities the other way around, that is, $P(T = \theta_j | \theta = \theta_k)$, and, as discussed, this is essentially what we did, though we didn't state it in this way originally. Now if we wanted to obtain the probabilities from (2.2) from this, we could try to make use of Bayes's formula, at least if we also assume or pretend here that θ itself is random:

$$P(\theta = \theta_j | T = \theta_j) = \frac{P(T = \theta_j | \theta = \theta_j) P(\theta = \theta_j)}{\sum_{k=1}^{N} P(T = \theta_j | \theta = \theta_k) P(\theta = \theta_k)} \tag{2.3}$$

However, we get stuck here, because we don't know the *a priori* probabilities $P(\theta = \theta_k)$. In fact, what we did looks reasonable precisely because we don't have access to the a priori probabilities. For example, if you somehow knew that, say, $P(\theta = \theta_1) = 0$, then it would

obviously be foolish to guess that $\theta = \theta_1$, even if this is what your carefully constructed statistic is suggesting.

To sum up this discussion, we should really be interested in the probabilities from (2.2), but these are out of reach, so what we actually do is only a compromise between the desirable and the feasible. The situation is similar in all of statistics: we typically control certain (conditional) probabilities, but if we could have it our way, we would be much more interested in those probabilities with the event and the condition switched.

If we make additional assumptions on the $P(\theta = \theta_k)$, then we can try to work with (2.3). In particular, if we assume that they are all equal to one another, then they will just drop out of (2.3), and we do obtain a formula for $P(\theta = \theta_j | T = \theta_j)$.

Exercise 2.4. You have a large supply of coins, all of them are biased, but half of them will come up heads more often, for the other half tails is more likely: let's say $\theta = 0.4$ or $\theta = 0.6$. You want to guess the θ of a given coin, based on a random sample of size $n = 5$. Proceed as described in the preceding paragraphs, and work out $P(\theta = 0.4 | T = 0.4)$.

Finally, let's look at a very specific example that will make the main point of this whole discussion in a vivid manner, I hope. We want to test people for psychic abilities by letting them predict the outcomes of coin tosses. (Tests are actually discussed in Section 6; we proceed informally here.) Someone without psychic abilities is reduced to guessing, which means that the probability of getting any given coin flip predicted correctly is 1/2. We apply rigorous standards and demand a *significance* of $\alpha = 0.001$. This means that someone is falsely declared a psychic only with probability $\alpha = 0.1\%$. (Or, in more technical language, the *null hypothesis* H_0 : *test subject does not have psychic abilities* is falsely rejected only with probability α when it is correct.) Since $(1/2)^{10} \simeq 0.001$, we can easily design such a test by letting the test subject predict 10 coin flips and declaring him or her a psychic if all 10 predictions were right.

Now the key question: We just declared A a psychic. What is the probability that this claim is correct? And how is this probability related to the significance? It is perfectly obvious in this example that the first question is unanswerable by statistical analysis only. It is about the real world. For example, if there are no psychics, then the probability of a correct claim in this scenario is zero, no matter how carefully the test was designed. Moving on to the second question, the significance has nothing to do with this; it controls a completely

different (conditional) probability, namely that of declaring A a psychic, given that A only guessed.

2.2. **Confidence intervals; sample variance.** Consider a random sample $X_1, \ldots, X_n$ drawn from a normal distribution $N(\theta, \sigma)$ with an unknown $\mu = \theta$; the standard deviation σ is assumed to be known to us. As we already discussed above, in a slightly different situation, if we wanted an estimator for θ, the obvious choice would be the sample mean $T = \overline{X} = (X_1 + \ldots + X_n)/n$.

Here, we would like to proceed somewhat differently. We are looking for a *confidence interval* $I = I(X_1, \ldots, X_n)$ that we expect will contain θ; or we could say we are looking for *two* statistics A, B, and now the interval $I = [A, B]$ will serve as our (somewhat ambiguous) guess on θ.

We would like to meet, to the extent we can, two somewhat contradictory requirements: we would like I to be small, and it should contain θ with reasonably high probability. It seems natural to try $I = [\overline{X} - d, \overline{X} + d]$ for suitable $d = d(X_1, \ldots, X_n) > 0$. By rearranging, we then have $\theta \in I$ precisely when

$$\theta - d \leq \overline{X} \leq \theta + d. \tag{2.4}$$

The probability of this event can be worked out; we'll make use of the following fact:

Theorem 2.3. *Suppose that $X_1 \sim N(\mu, \sigma)$. Then the sample mean $\overline{X}$ of a random sample is $N(\mu, \sigma/\sqrt{n})$-distributed.*

Proof. It suffices to establish the following slightly more general claim (as far as the distributions are concerned) for a sample of size $n = 2$: Suppose that X, Y are independent and normally distributed, with possibly distinct parameters. Then $Z = X + Y \sim N(\mu_X + \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2})$. This we will verify by a brute force calculation.

Exercise 2.5. Show how the claim of Theorem 2.3 can then be obtained by repeatedly applying this.

By independence, the joint density of X, Y is the product of the individual densities.

Exercise 2.6. Verify this. More specifically, check that if we define

$$Q_{X,Y}(A \times B) = \iint_{A \times B} f_X(x) f_Y(y)\, dxdy,$$

then $Q(A \times B) = P(X \in A)P(Y \in B)$. (This implies that $Q(C) = P((X, Y) \in C)$ for $C \subseteq \mathbb{R}^2$, as required.)

Then

$$P(Z\le z)= \frac{1}{2\pi\sigma_X\sigma_Y}\int_{-\infty}^{\infty}dx\int_{-\infty}^{z-x}dy\,\exp\left(-\frac{(x-\mu_X)^2}{2\sigma_X^2}-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}\right), \tag{2.5}$$

so, since $f_Z(z)=(d/dz)P(Z\le z)$, we obtain

$$f_Z(z)=\frac{1}{2\pi\sigma_X\sigma_Y}\int_{-\infty}^{\infty}\exp\left(-\frac{(x-\mu_X)^2}{2\sigma_X^2}-\frac{(z-x-\mu_Y)^2}{2\sigma_Y^2}\right)dx. \tag{2.6}$$

The argument of the exponential function can be rewritten as follows: it equals

$$-\frac{1}{2}\left(\frac{1}{\sigma_X^2}+\frac{1}{\sigma_Y^2}\right)\left[x^2-2\frac{\mu_X\sigma_Y^2+(z-\mu_Y)\sigma_X^2}{\sigma_X^2+\sigma_Y^2}\,x + \frac{\mu_X^2\sigma_Y^2}{\sigma_X^2+\sigma_Y^2}+\frac{(z-\mu_Y)^2\sigma_X^2}{\sigma_X^2+\sigma_Y^2}\right].$$

The first two terms in square brackets have the form x^2-2Ax. We complete the square. Then the integrand from (2.6) takes the form $\exp(-(x-A)^2/(2B^2))\exp(\ldots)$, where the second exponent is independent of x and thus just acts as a factor as far as the integration is concerned. If we keep track of things more carefully, we find that

$$A=\frac{\mu_X\sigma_Y^2+(z-\mu_Y)\sigma_X^2}{\sigma^2},\quad B=\frac{\sigma_X\sigma_Y}{\sigma},\quad \sigma^2\equiv\sigma_X^2+\sigma_Y^2. \tag{2.7}$$

In particular, what we are integrating is a normal density again, so we know the value of the integral: $\int\exp(-(x-A)^2/(2B^2))\,dx=\sqrt{2\pi}B$. If we make use of this, then (2.6) becomes

$$f_Z(z)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2\sigma_X^2\sigma_Y^2}\cdot\text{Exponent}\right), \tag{2.8}$$

where

$$\text{Exponent}=-(\sigma^2A)^2+\mu_X^2\sigma_Y^2\sigma^2+(z-\mu_Y)^2\sigma_X^2\sigma^2.$$

This simplifies further if we plug in A from (2.7) and multiply out the square:

$$\begin{aligned}\text{Exponent}&=\sigma_X^2\sigma_Y^2(\mu_X^2+(z-\mu_Y)^2-2\mu_X(z-\mu_Y))\\&=\sigma_X^2\sigma_Y^2(z-\mu_X-\mu_Y)^2\end{aligned}$$

So f_Z from (2.8) is indeed the density of a normal distribution with parameters $\mu_X+\mu_Y$ and σ, as claimed. □

Exercise 2.7. The following fact was tacitly used in this argument: If $X \sim N(\mu, \sigma)$ and $a > 0$, then $aX \sim N(a\mu, a\sigma)$. Verify this.

Exercise 2.8. Show that the full claim about Z follows as soon as we can show that $f_Z(z) = Ce^{-D(z-E)^2}$, for *some* constants $C, D > 0$, $E \in \mathbb{R}$. (This could have been used to try to save some work in the part after (2.8).)

Exercise 2.9. Use the calculation from the first part of the proof of Theorem 2.3 to establish the following general fact: if X, Y are independent continuous random variables, then $Z = X + Y$ has density

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(t) f_Y(z-t)\, dt. \tag{2.9}$$

Exercise 2.10. The RHS of (2.9) is called the *convolution* of f_X, f_Y and is also denoted by $f_X * f_Y$. Show that $f * g = g * f$. (Do this honestly, by a calculation; however, why is it in fact already clear, from the result of the previous Exercise, that this has to be true?)

Exercise 2.11. Recall that the *moment generating function* of a random variable X is defined as $M_X(t) = Ee^{tX}$, provided this expectation converges.
(a) Compute $M_X(t)$ for an $N(\mu, \sigma)$-distributed X. Show that

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right).$$

(b) Use this to give a much slicker (and much quicker) proof of Theorem 2.3.

We finally return to (2.4). We now know that $\overline{X} \sim N(\theta, \sigma/\sqrt{n})$, so

$$\frac{\sqrt{n}}{\sigma}(\overline{X} - \theta) \sim N(0,1) \tag{2.10}$$

(compare Exercise 2.7). An $N(0,1)$-distributed random variable will be within 2 standard deviations ($= 2$) of its expectation ($= 0$) with probability $\simeq 0.954$; more generally, for any α, we can extract a z_α from a table so that

$$P(-z_\alpha \leq Z \leq z_\alpha) = 1 - \alpha. \tag{2.11}$$

So the event

$$-2\frac{\sigma}{\sqrt{n}} \leq \theta - \overline{X} \leq 2\frac{\sigma}{\sqrt{n}}$$

has probability 0.954, and we have achieved what we set out to do: $I = [\overline{X} - 2\sigma/\sqrt{n}, \overline{X} + 2\sigma/\sqrt{n}]$ is a 0.954 confidence interval for θ. We summarize:

Theorem 2.4. *Suppose that $\overline{X}$ is the sample mean of a random sample of size n of $N(\theta,\sigma)$-distributed random variables. For a significance $\alpha > 0$, typically small, let z_α be as in (2.11), where $Z \sim N(0,1)$. Then the (random) interval*

$$I = \left[\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}}, \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right] \tag{2.12}$$

is a $(1-\alpha)$ confidence interval for θ: $P(\theta \in I) = 1-\alpha$

Note also that while in principle the discussion from the end of the previous section applies here too, the situation is quite satisfactory now: we compute an interval, let's say at significance $\alpha = 0.05$, and if this whole procedure is repeated many times, then our claim that the interval produced contains θ will be right 95% of the time. In more mathematical terms, what works to our advantage is the fact that $\overline{X} - \theta \sim N(0, \sigma/\sqrt{n})$, *independently of* θ.

Frequently, the random variables we are interested in will not be normally distributed, and then Theorem 2.4 as stated doesn't apply. We can try to work around this, however. First of all, the Central Limit Theorem guarantees that for any distribution, the random variable from (2.10) will at least be approximately $N(0,1)$-distributed for large n, provided that, as before, $\theta = EX_1$, $\sigma^2 = \mathrm{Var}(X_1)$.

Another difficulty is that σ will usually not be known to us. We solve this problem by also using the data to estimate σ. More specifically, we make use of the *sample variance*

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2.$$

Notice that this is a statistic. Apart from the slightly unexpected $n-1$ (instead of n) in the denominator, this looks like the obvious thing to try; the formula just mimics the definition of the actual variance, making do with what we have available. Since we will mainly be interested in large sample sizes n, it doesn't make a big difference whether we use $n-1$ or n in the denominator, so we shouldn't get very upset about this, and anyway the choice is justified by the following theorem.

Theorem 2.5. *For any distribution, S^2 is an unbiased estimator for $\sigma^2 = \mathrm{Var}(X_1)$.*

Exercise 2.12. Toss a fair coin twice, that is, $P(X_1 = 0) = P(X_1 = 1) = 1/2$, $n = 2$.
(a) Confirm that indeed $ES^2 = \mathrm{Var}(X_1)$. You can do this entirely by hand; just set up a sample space with four points in it.

(b) However, show that $ES \neq \sigma_{X_1}$, so S is not an unbiased estimator of the standard deviation.

In view of this, shouldn't we be looking for an unbiased estimator of the *standard deviation,* rather than one of the *variance?* Perhaps so, but mathematical convenience trumps all other considerations here; S^2 as defined above is easy to work with.

Proof. We compute

$$\sum_{j=1}^{n}(X_j - \overline{X})^2 = \sum_{j=1}^{n}(X_j^2 + \overline{X}^2 - 2X_j\overline{X}) = \sum_{j=1}^{n}(X_j^2 - \overline{X}^2),$$

since $\sum X_j\overline{X} = n\overline{X}^2 = \sum \overline{X}^2$. Thus

$$(n-1)ES^2 = n(EX_1^2 - E\overline{X}^2), \tag{2.13}$$

and we now evaluate $E\overline{X}^2$ by a similar calculation:

$$\begin{aligned} E\overline{X}^2 &= \frac{1}{n^2}\sum_{j,k=1}^{n} EX_jX_k = \frac{1}{n^2}\left(\sum_{j=1}^{n} EX_j^2 + \sum_{j\neq k} EX_jEX_k\right) \\ &= \frac{EX_1^2}{n} - \frac{n^2-n}{n^2}(EX_1)^2 \end{aligned}$$

Here (and only here) we have used independence to evaluate EX_jX_k. If we plug this into (2.13), then we find that

$$(n-1)ES^2 = n\left(1-\frac{1}{n}\right)\left(EX_1^2 - (EX_1)^2\right) = (n-1)\mathrm{Var}(X_1),$$

as desired. □

So for an arbitrary distribution, we can now proceed as follows to set up an approximate confidence interval for the unknown mean $\theta = EX_1$, which (we hope) will be valid for large n. Let $x_1, \ldots, x_n$ denote the observed data: (1) work out the sample standard deviation $S = S(x_1, \ldots, x_n)$; (2) use (2.12), with σ replaced by S.

For the following problems, you will need a table of (the cumulative distribution function of) the standard normal distribution. If you don't have one available, see:

`http://en.wikipedia.org/wiki/Standard_normal_table`

Exercise 2.13. Suppose that for an $N(\theta, 3)$-distributed random variable, the sample mean $\overline{X} = 10$ was observed for a random sample of size $n = 20$. Find a 95% confidence interval for θ.

Exercise 2.14. You would like to obtain a 90% confidence interval for the unknown mean θ of an $N(\theta, 3)$-distribution of length at most 2. In other words, you would like to use $I = [\overline{X} - 1, \overline{X} + 1]$. What is the minimal sample size consistent with these requirements?

Exercise 2.15. Toss a coin $n = 10,000$ times; the probability of heads $\theta = P(X_1 = 1)$ is unknown. You find that heads occurred $6,000$ times.
(a) Use the method suggested above (replace σ by the sample standard deviation) to find an approximate 99% confidence interval for θ.
(b) Alternatively, play it by ear, as follows: recall the formula $\sigma = \sqrt{\theta(1-\theta)}$, and use $\overline{X}$ as an estimator for θ to work around the fact that θ is not known. Compare with the result from part (a).
(c) Same as (a), but now assume that heads occurred $9,000$ times.

Let's now return to the situation where the members of the random sample $X_1, \ldots, X_n$ are specifically $N(\mu, \sigma)$-distributed, but with an unknown σ. We just said that then we'll use

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

and pretend (backed up by the CLT) that $T \sim N(0,1)$. This will work just fine for large n, but of course it is not justified for small n.

It turns out that we can actually determine the *exact* distribution of T. We'll discuss this in some detail; this will also give us the opportunity to introduce two new distributions that come up quite regularly in statistics.

To get this started, suppose that $X \sim N(0,1)$. What is the distribution of $Y = X^2$? This is a routine calculation. Clearly,

$$P(Y \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = 1 - 2P(X > \sqrt{y}) =$$
$$= 1 - \sqrt{\frac{2}{\pi}} \int_{\sqrt{y}}^{\infty} e^{-t^2/2}\, dt,$$

and thus

$$f_Y(y) = \frac{d}{dy} P(Y \le y) = \frac{e^{-y/2}}{\sqrt{2\pi y}}.$$

More precisely, this formula is valid for $y \ge 0$, and obviously $f_Y(y) = 0$ for $y < 0$.

Definition 2.6. We say that a random variable X is $\chi^2(n)$-distributed if $X \ge 0$ and has the density

$$f(x) = c_n x^{n/2-1} e^{-x/2} \qquad (x \ge 0).$$

Here, the constant $c_n > 0$ is chosen so that $\int_0^\infty f_n(x)\,dx = 1$. In this situation, we also call X χ^2-distributed with n *degrees of freedom.*

Exercise 2.16. The *gamma function* is defined as (for $x > 0$)

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt.$$

Show that $c_n = 1/(2^{n/2}\Gamma(n/2))$.

So what we just did can now be rephrased as follows: if $X \sim N(0,1)$, then $X^2 \sim \chi^2(1)$. Now let's see if we can also handle a sum $X_1^2 + \ldots + X_n^2$, with X_j iid $N(0,1)$-distributed random variables.

Theorem 2.7. *Suppose that $Y_1, \ldots, Y_k$ are independent and $Y_j \sim \chi^2(n_j)$. Then $Y_1 + \ldots + Y_k \sim \chi^2(n)$, with $n = n_1 + \ldots + n_k$.*

Proof. This is similar to Theorem 2.3, but it will be considerably less tedious, since we've now learnt our lesson from that calculation. Also, I'll only do the case $k = 2$, as an illustration. From Exercise 2.9, we know that the sum of independent random variables has a density equal to the convolution of the individual densities. To make this more concrete, if $Y_1 \sim \chi^2(m)$, $Y_2 \sim \chi^2(n)$, then the density f of the sum is given by

$$f(x) = \int_{-\infty}^{\infty} f_{Y_1}(t) f_{Y_2}(x-t)\,dt = c_m c_n e^{-x/2} \int_0^x t^{m/2-1}(x-t)^{n/2-1}\,dt.$$

This integral still looks moderately scary, but fortunately we only need to know its dependence on x, so we can proceed as follows:

$$\begin{aligned}\int_0^x t^{m/2-1}(x-t)^{n/2-1}\,dt &= x^{m/2+n/2-2}\int_0^x (t/x)^{m/2-1}(1-t/x)^{n/2-1}\,dt \\ &= x^{m/2+n/2-1}\int_0^1 s^{m/2-1}(1-s)^{n/2-1}\,ds\end{aligned}$$

That worked beautifully! The integral is constant, as a function of x; it only depends on m, n. So we conclude that

$$f(x) = C(m,n)x^{m/2+n/2-1}e^{-x/2}.$$

We know that f is a density, and we now see that it has the same form as the density of the χ^2 distribution with $m+n$ degrees of freedom, except possibly for the constant. This however implies that $C(m,n) = c_{m+n}$, as desired, because this is the constant that makes $\int f = 1$. □

Now let's return to an $N(\mu,\sigma)$-distributed random sample and let's see what this says about S^2. Observe that

$$\begin{aligned}(n-1)S^2 &= \sum (X_j-\overline{X})^2 = \sum (X_j-\mu-(\overline{X}-\mu))^2\\ &= \sum (X_j-\mu)^2 - 2(\overline{X}-\mu)\sum (X_j-\mu) + n(\overline{X}-\mu)^2\\ &= \sum (X_j-\mu)^2 - n(\overline{X}-\mu)^2.\end{aligned}$$

In other words,

$$\frac{n-1}{\sigma^2}S^2 = \sum \left(\frac{X_j-\mu}{\sigma}\right)^2 - \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2. \tag{2.14}$$

By Theorem 2.7, the sum is $\chi^2(n)$-distributed, and since $\sqrt{n}(\overline{X}-\mu)/\sigma \sim N(0,1)$, the formula seems to suggest that $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$. This is indeed correct.

Exercise 2.17. Check more carefully that

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Theorem 2.8. *Let $X_1,\dots,X_n$ be a random sample drawn from an $N(\mu,\sigma)$ distribution. Then:*
(a) $\overline{X} \sim N(\mu,\sigma/\sqrt{n})$;
(b) $\overline{X}$, S^2 are independent;
(c) $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$;
(d) $T=(\overline{X}-\mu)/(S/\sqrt{n})$ is t_{n-1}-distributed.

Part (d) introduces a new distribution, the *Student* (after the *nom de plume* of William Gosset, who used this distribution in an article in 1908) or *t distribution:* We say that a random variable T is t_m-distributed (and we again refer to m as *degrees of freedom*) if T has density

$$f(t) = \frac{d_m}{(1+t^2/m)^{(m+1)/2}};$$

the normalization constant $d_m = \Gamma((m+1)/2)/(\sqrt{\pi m}\Gamma(m/2))$ can again be expressed in terms of Γ functions. The general fact about this distribution that is used to obtain part (d) of the theorem is that if U,V are independent and $U\sim N(0,1)$, $V\sim\chi^2(m)$, then $U/\sqrt{V/m}\sim t_m$.

Exercise 2.18. Derive the density f of the Student distribution from this description.

I don't want to prove Theorem 2.8 in detail here. Once we have (b), the rest pretty much falls into place. Part (a) is essentially Theorem 2.3

and was established again in Exercise 2.17 above. Part (c) follows from (b) and (2.14), more or less as discussed above. Part (d) also follows from (b) and the properties of the Student distribution, as discussed in the previous Exercise. (We will actually see a very elegant proof of (b), which depends on more advanced statistical tools, much later, in Section 5.3.)

Let's now return to the example that motivated this whole discussion. We have a random sample drawn from an $N(\mu, \sigma)$ distribution, with μ and σ both unknown, and we would like to build a $(1 - \alpha)$ confidence interval for μ. Say $\alpha = 0.05$, to make this concrete. Then we define $t = t_{0.05,n}$ by the property that a t_{n-1}-distributed random variable T satisfies $P(-t \leq T \leq t) = 0.95$. With this definition of t, Theorem 2.8(d) will then show that

$$\overline{X} - t\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t\frac{S}{\sqrt{n}}$$

with probability 0.95. In other words, $[\overline{X} - tS/\sqrt{n}, \overline{X} - tS/\sqrt{n}]$ is the exact 0.95 confidence interval we've been looking for. We in principle find t from the condition that

$$d_{n-1} \int_{-t}^{t} \frac{dx}{(1 + x^2/(n-1))^{n/2}} = 0.95,$$

but of course it is easier to just extract these values from a table. For example, $t_{0.05,5} \simeq 2.78$, $t_{0.05,10} \simeq 2.26$, $t_{0.05,20} \simeq 2.09$, $t_{0.05,30} \simeq 2.05$.

If we had just used the normal approximation discussed earlier, then instead of t, we would be using the corresponding value z for the standard normal distribution, that is,

$$\frac{1}{\sqrt{2\pi}} \int_{-z}^{z} e^{-x^2/2}\, dx = 0.95.$$

This equals $z \simeq 1.96$, so, as expected, we obtain a reasonable approximation for n not too small.

Exercise 2.19. Assume that the score of a randomly chosen student on a given test is normally distributed, with unknown μ, σ (this is obviously not literally correct; for example, the possible scores get neither negative nor arbitrarily large). In a class with $n = 5$ students, the scores were $1, 48, 50, 62, 99$. Use the t distribution to find a 95% confidence interval for μ.

Exercise 2.20. Suppose that T is t_n-distributed, with $n \geq 2$.
(a) Use a direct computation to find ET and $\mathrm{Var}(T)$ (hint: use integration by parts to compute ET^2; express this in terms of the constants d_n).

(b) Use the fact (mentioned above, see Exercise 2.17) that if $U \sim N(0,1)$, $V \sim \chi^2(n)$, U, V independent, then $U/\sqrt{V/n} \sim t_n$ to derive these answers in a different way.