The Probability Lifesaver: Order Statistics and the Median Theorem

Steven J. Miller

December 30, 2015

Contents

Greetings again!

In this supplemental chapter we develop the theory of order statistics in order to prove The Median Theorem. This is a beautiful result in its own, but also extremely important as a substitute for the Central Limit Theorem, and allows us to say non-trivial things when the CLT is unavailable.

Chapter 1

Order Statistics and the Median Theorem

The Central Limit Theorem is one of the gems of probability. It's easy to use and its hypotheses are satisfied in a wealth of problems. Many courses build towards a proof of this beautiful and powerful result, as it truly is 'central' to the entire subject.

Not to detract from the majesty of this wonderful result, however, what happens in those instances where it's unavailable? For example, one of the key assumptions that must be met is that our random variables need to have finite higher moments, or at the very least a finite variance. What if we were to consider sums of Cauchy random variables? Is there anything we can say?

This is not just a question of theoretical interest, of mathematicians generalizing for the sake of generalization. The following example from economics highlights why this chapter is more than just of theoretical interest. For years many financial models assumed that price changes were drawn from normal random variables; however, Mandelbrot (of fractal geometry fame) and others have argued that these models are inconsistent with real world data. They were led to this belief by looking at large fluctuation days; the number of such days with very large changes was magnitudes higher than normal theory predicts. In other words, the probability of observing as many days with such high percentage drops in the major stock markets under the assumption that price changes come from normal distributions is so small that this hypothesis must be discarded. In its place they suggest a complicated model, which at its core has Cauchy distributions and infinite variance.

Thus, if we (or our friends or colleagues) care about economics, it might be more than just an academic question as to what replaces the Central Limit Theorem when the variances are infinite. The replacement involves, not the mean, but the median.

The **median** $\widetilde{\mu}$ of a random variable X are all points x such that

$$\text{Prob}(X \le x) \ \ge\ \frac{1}{2} \quad \text{and} \quad \text{Prob}(X \ge x) \ \ge\ \frac{1}{2}.$$

If our density is continuous, any point x such that half the probability occurs before x and half occurs after x is a median. We'll discuss this definition in much greater detail below.

There is much that can be said and proved, at least in the special case when our random variables are symmetric about some point (typically the probabilities will be symmetric about the origin, so that the probability density p satisfies $p(x) = p(-x)$). For such distributions the median is the same as the mean. Thus instead of looking at the sample mean we can study the sample median; the advantage is that there are situations where the sample median converges to a nice distribution while the sample mean does not.

The exposition gets a bit involved; to simplify the descriptions, it's convenient to use big-Oh notation. We describe this in great detail in **ADD REF**. If you haven't seen this notation before, you should spend some time reading **ADD REF** now; if you have seen this before, the quick summary below should suffice.

Definition 1.0.1 (Big-Oh Notation) *$A(x) = O(B(x))$, read "$A(x)$ is of order (or big-Oh) $B(x)$", means there is a $C > 0$ and an x_0 such that for all $x \ge x_0$, $|A(x)| \le C\,B(x)$. This is also written $A(x) \ll B(x)$ or $B(x) \gg A(x)$.*

Big-Oh notation is a convenient way to handle lower order terms. For example, if we write $F(x) = x^5 + O(x^2)$, this means that as x tends to infinity, the main term of $F(x)$ grows like x^5, and the correction (or error) terms are at most some constant times x^2. Useful examples include for $r, \epsilon > 0$, as $x \to \infty$ we have $x^r = O(e^x)$ and $\log x = O(x^\epsilon)$.

The main result is:

Median Theorem: Let a sample of size $n = 2m + 1$ with n large be taken from an infinite population with a density function $f(\tilde{x})$ that is nonzero at the population median $\tilde{\mu}$ and continuously differentiable in a neighborhood of $\tilde{\mu}$. The sampling distribution of the median is approximately normal with mean $\tilde{\mu}$ and variance $\frac{1}{8f(\tilde{\mu})^2 m}$.

In many cases, this theorem is a good substitute for the Central Limit Theorem. Note the commonalities between the two. Both involve an average of sample values, and both have a sample statistic converging to a normal distribution. There are, however, major differences between the two. The most glaring, of course, is that we're looking at different quantities (the mean versus the median). The next is that the Median Theorem allows the density to have infinite variance. Finally, it's worth noting that the Central Limit Theorem requires us to standardize our sample statistic; no such standardization is required here.

Before developing the analogue of the Central Limit Theorem for the median, we study a subject of interest in its own right, order statistics. These will be essential in our analysis of the median. In this chapter we'll use results on the multinomial distribution, Taylor series expansions, and Stirling's formula; you should consult those sections of the book if you have any questions on these.

1.1 Definition of the Median

Key points for this section:

1. The definition of the median.
2. Examples of finding the median.
3. Applications to estimating means.

The definition of the median is one of the most cumbersome statements you'll find about a concept that seems obvious. It's quite unfortunate, as the informal definition, which works in many cases, sadly fails in some cases; this failure requires us to have the wordy monstrosity below.

The quick and painless definition of the median is the point where half the probability comes before it and half comes after it. As mentioned above, this definition isn't quite right, but it does express the general idea quickly. As always, we'll consider our densities as either being discrete or continuous. We'll first look at some examples and use those to get a better sense as to what the right definition of median should be.

Let's study a discrete distribution first. Imagine we have a fair die. Then the possible outcomes are 1, 2, 3, 4, 5 and 6, and each occurs with probability 1/6. We want to find a value $\widetilde{\mu}$ such that half the probability occurs before $\widetilde{\mu}$ and half occurs afterwards. Looking at a plot of the cumulative distribution function (which we give in Figure 1.1), we see that *any* number in $[3, 4)$ works; note we include the left endpoint but not the right endpoint.

Why? Let's take x to be a positive real number. Then

$$\text{Prob(die is at most } x) \;=\; \sum_{\substack{n=1\\ n\le x}}^{6} \frac{1}{6}.$$

This means that for any $x \in [3, 4)$, the probability of being at most x is just the probability of being either a 1, 2 or 3; this is just $1/6 + 1/6 + 1/6$ or 1/2.

There are a lot of important things to learn from the above example. The first is that the median isn't always unique; there could be a range of values. Second, the medians don't have to generate an open interval or a closed interval; our example

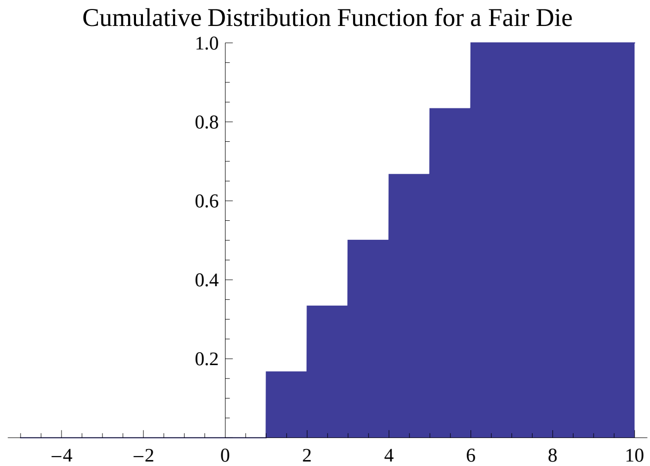


Figure 1.1: Plot of the cumulative distribution function (CDF) for a fair die. Note the CDF is zero for $x < 1$, is a nice step function for $1 \leq x < 7$, and is identically 1 for $x \geq 7$.

above is an open interval.

Let's look at a more complicated example. We'll assume the only possible outcomes are 1, 2, 3, 4, 5 and 6, and that $\text{Prob}(X = n) = n/21$. This is clearly a probability density, as the values are all non-negative and sum to 1. We may interpret this by viewing this as a study of a weighted die. We plot the cumulative distribution function in Figure 1.2.

We see

$$\text{Prob}(X < 5) \;=\; \frac{1}{21} + \frac{2}{21} + \frac{3}{21} + \frac{4}{21} \;=\; \frac{10}{21} \;\approx\; 0.47619.$$

Thus no value less than 5 can be the median, as the probability of being *strictly* less than 5 is less than 50%. Similarly, we see

$$\text{Prob}(X > 5) \;=\; \frac{6}{21} \;\approx\; 0.285714.$$

Thus no value 5 or larger can be the median! Only one possibility is left, namely that 5 is the median, and the median is unique in this problem. In this case, we have

$$\text{Prob}(X \leq 5) \;=\; \frac{15}{21} \;\approx\; 0.714286 \quad \text{and} \quad \text{Prob}(X \geq 5) \;=\; \frac{11}{21} \;\approx\; 0.52381.$$

Notice how different the answer is in the second example to the first. With the weighted die, there is a unique median. The lesson to be gleaned from all of this is that, unfortunately, what the median is can depend wildly on the distribution. Sometimes there is a unique median; sometimes there is an interval of medians.

For the continuous distributions that we meet in a typical probability course, the situation is fortunately much nicer. In fact, as long as the density never vanishes

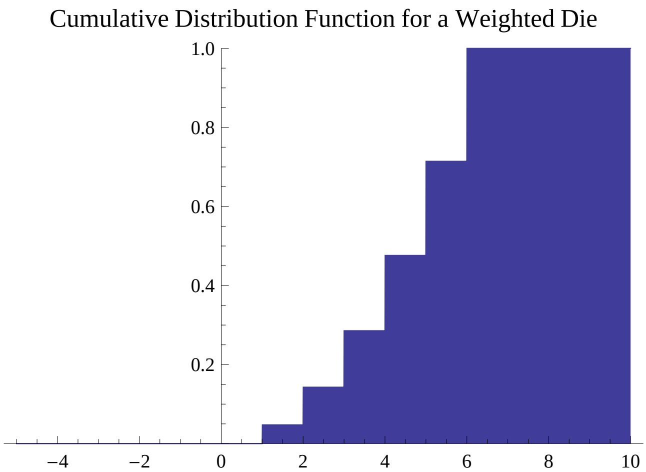


Figure 1.2: Plot of the cumulative distribution function (CDF) for a weighted die, where the probability of rolling an n is $n/21$. Note the CDF is zero for $x < 1$, is a nice step function for $1 \leq x < 7$, and is identically 1 for $x \geq 7$.

except on an interval of the form $(-\infty, a)$, $(-\infty, a]$, $[b, \infty)$ or (b, ∞), there is a unique median! Unlike the discrete case, for continuous densities the cumulative distribution function is continuous. This means there are no jumps, and it's the jumps that caused all the headaches above!

Let's take one of the standard favorites, the exponential distribution with density

$$f(x) \;=\; \begin{cases} e^{-x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

The cumulative distribution is easily calculated; it's

$$F(x) \;=\; \int_{-\infty}^{x} f(t)dt \;=\; \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - e^{-x} & \text{if } x \geq 0. \end{cases}$$

The median is the value $\widetilde{\mu}$ such that $F(\widetilde{\mu}) = 1/2$. For this density, that means

$$1 - e^{-\widetilde{\mu}} \;=\; \frac{1}{2}.$$

which after some algebra is seen to be $\widetilde{\mu} = \log 2$.

We end with one last example to show that things be annoying for continuous random variables as well. Consider the density depicted in Figure 1.3, with cumulative distribution function given in Figure 1.4. Note that there is no unique median; any number in $[-1, 1]$ is a median. This degeneracy is due to the fact that there is *no* probability of taking on a value in $[-1, 1]$, and thus the cumulative distribution function is flat from -1 to 1.

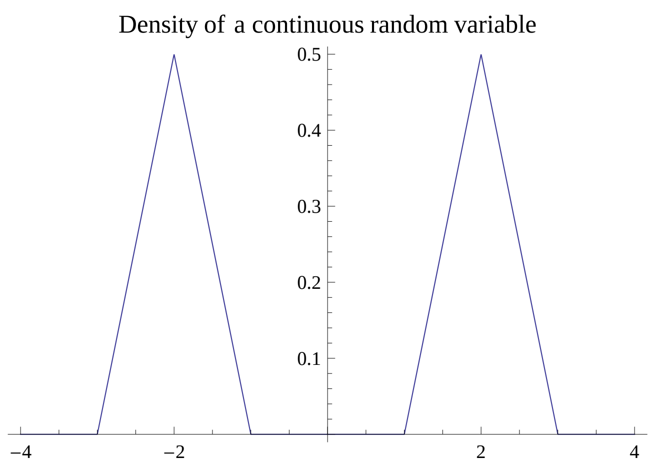


Figure 1.3: Plot of the density function $f(x) = \frac{1}{4} - \frac{1}{4}\,||x| - 2|$ for $1 \leq |x| \leq 3$ and 0 otherwise.

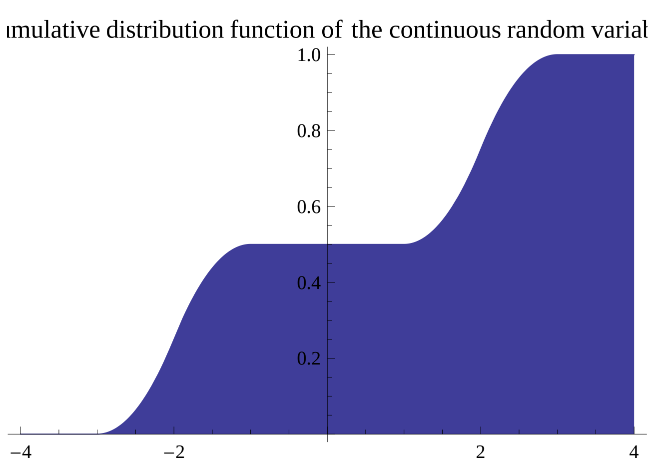


Figure 1.4: Plot of the cumulative distribution function for the density $f(x) = \frac{1}{4} - \frac{1}{4}\,||x| - 2|$ for $1 \leq |x| \leq 3$ and 0 otherwise.

To sum up, even though we would like to think of the median as that point where half the probability falls before and half falls after, this definition isn't quite right. For most distributions, either there is no unique median, or there is a median but the probabilities are not 50%.

We end with a very important observation:

> Let X be a random variable with density p that is symmetric about its mean μ; this means $p(\mu - t) = p(\mu + t)$. Then the median $\widetilde{\mu}$ equals the mean μ.

The utility of this point of view will become more and more apparent as we proceed through the chapter. For distributions that are symmetric about the mean, if we can estimate the median then we get for free an estimate of the mean!

As always, whenever we're looking for an example of a probability density that has a mean but no variance, we should immediately think of a Cauchy distribution (or a cousin). For example, consider the density

$$f(x) \;=\; \frac{1}{\pi}\frac{1}{1+(x-\mu)^2}.$$

It is debateable as to whether or not this distribution has a mean. If it were to have one, it would clearly be μ. Why might it not have a mean? The problem is what does

$$\int_{-\infty}^{\infty} x f(x) dx$$

mean? Everything is fine if we interpret the integral as

$$\lim_{A\to\infty} \int_{-A}^{A} x f(x) dx,$$

as in this case it's a finite integral of an odd function, and thus zero; if however we interpret it as

$$\lim_{A,B\to\infty} \int_{-A}^{B} x f(x) dx$$

then the answer depends on how A and B tend to infinity. For example, if $B = 2A$ then the integral is

$$\lim_{A\to\infty} \int_{A}^{2A} \frac{x}{\pi(1+(x-\mu)^2)} dx.$$

For x large, the integrand is essentially $1/\pi x$, and thus the integral diverges. Thus, annoying as it is to say it, the Cauchy distribution might not even have a mean; however, it does have a median, and the median is readily seen to be μ. Thus the median can serve as a surrogate to the mean.

If the previous example felt a bit unsatisfying as the mean didn't even exist, consider instead

$$g(x) \;=\; \frac{3\sqrt{3}}{4\pi}\frac{1}{1+|x-\mu|^3}.$$

It turns out g is a probability density. While it's clearly non-zero, it's not at all apparent that it integrates to 1. The easiest way to see this is to use Cauchy's residue theorem from Complex Analysis, but this is heavy machinery. Fortunately, it's clear that if we consider

$$h(x) \;=\; \frac{C}{1+|x-\mu|^3}$$

then there is some choice of C such that this integrates to 1, namely

$$C \;=\; \left(\int_{-\infty}^{\infty} \frac{1}{1+|x-\mu|^3}dx\right)^{-1}.$$

There is some constant that works; the actual value is immaterial. Now the mean clearly exists, as its defining integral is

$$\int_{-\infty}^{\infty} \frac{3\sqrt{3}}{4\pi}\frac{x}{1+|x-\mu|^3}dx;$$

the integrand decays like $1/x^2$ and thus there are no convergence issues. Unfortunately this distribution has infinite variance, which means the Central Limit Theorem won't be available to estimate its mean. We'll see below, however, that we *can* estimate the median, and as the median and the mean are equal, we can use the median estimate for the mean.

1.2 Order Statistics

Key points for this section:

1. The definition of order statistics.
2. Finding the densities of order statistics.

Suppose that the random variables $X_1, X_2, \dots, X_n$ constitute a sample of size n from an infinite population with continuous density. Often it's useful to reorder these random variables from smallest to largest. In reordering the variables, we rename them so that Y_1 is a random variable whose value is the smallest of the X_i, Y_2 is the next smallest, and so on, with Y_n the largest of the X_i. Y_r is called the r^{th} **order statistic** of the sample.

We write this more explicitly as

$$Y_1 \;=\; \min_i(X_1,\dots,X_n),\;\dots,\; Y_n \;=\; \max_i(X_1,\dots,X_n).$$

If we had

$$X_1 \;=\; 5,\; X_2 \;=\; 23,\; X_3 \;=\; 1,\; X_4 \;=\; 44,\; X_5 \;=\; 6,$$

then

$$Y_1 \;=\; 1,\; Y_2 \;=\; 5,\; Y_3 \;=\; 6,\; Y_4 \;=\; 23,\; Y_5 \;=\; 44.$$

There are many reasons we might want to know all we can about these order statistics. For three of these the applications are fairly obvious, namely for Y_1 (the smallest value), Y_n (the largest value), and the median. If n is odd the median is just $Y_{n/2+1}$, while if n is even the median is any value in $[Y_{n/2}, Y_{n/2+1}]$. We'll give just one real world reason. Imagine we're building an elevator or a bridge, and our random variables are the weight on the bridge at various times during the day. We should be very concerned with how much weight it can bear. This would thus be a situation where we care about the distribution of the maximum.

OK. You're hopefully now somewhat convinced that we care about these order statistics. How do we find them? It turns out that we can derive an expression for the probability density of the r^{th} order statistic.

Theorem 1.2.1 *For a random sample of size n from an infinite population having values x and continuous density $f(x)$, the probability density of the r^{th} order statistic Y_r is given by*

$$g_r(y_r) \;=\; \frac{n!}{(r-1)!(n-r)!}\left[\int_{-\infty}^{y_r} f(x)\,dx\right]^{r-1} f(y_r)\left[\int_{y_r}^{\infty} f(x)\,dx\right]^{n-r}.$$

The theorem holds under weaker conditions than stated above; we can replace continuity with many things, such as the density being bounded. We'll phrase it as we have above as this is general enough to cover many cases of interest, and leads to slightly easier proofs.

As always, our first thought after encountering such a long and involved formula is to explore some special cases. The simplest is to imaging that $n = 1$. If this happens, then there is only one possible order statistic, namely $r = 1$, and the density collapses to $f(y_1)$. This follows from some algebra, using $0! = 1$ and anything raised to the zeroth power is 1.

We should look for some other simple tests, as the one above feels like cheating; we're not really looking at order statistics if n is just 1! Thus, let's consider $n \geq 2$ and compare the densities of Y_1 and Y_n. There are a lot of commonalities between the two. In each case the ratio of factorials equals n, both have a lone evaluation of f, and both have just one of the integrals raised to a non-zero power. To highlight the commonalities and differences, we use the same dummy variable for each, and find

$$\begin{aligned} f_1(y) &= nf(y)\left(\int_y^{\infty} f(x)dx\right)^{n-1} \\ f_n(y) &= nf(y)\left(\int_{-\infty}^{y} f(x)dx\right)^{n-1}. \end{aligned}$$

Let's look at this and try and get a feel for what it's saying. In situations like this, it's a good idea to go for the extremes; imagine y is near infinity near negative infinity. If y is very large, then $\int_y^{\infty} f(x)dx$ is very small, while $\int_{-\infty}^{y} f(x)dx$ is very close to 1. Thus, for y large, $f_1(y)$ is much smaller than $f_n(y)$. If instead we looked at y near negative infinity, we would find $f_1(y)$ is much larger than $f_n(y)$. This makes perfect sense! Clearly $Y_1 \leq Y_n$, as the smallest value cannot exceed the largest! Thus we expect more of the mass of Y_1 to occur for 'small' y than Y_n, and similarly for large y we expect Y_n to have more mass.

It's good advice to always try quick arguments like this whenever you see a theorem. Test it. Look at special cases. This is a great way to get a feel for what it's saying. Only after playing with the formula should you turn to the proof. Speaking of the proof, it's sadly a little long and technical in places. It's fine to just skim it (or even skip it). We chose to include the proof as it highlights a lot of the powerful methods used to prove probability results in the field.

Proof of Theorem 1.2.1: Let h be a positive real number. We divide the real line into three intervals: $(-\infty, y_r)$, $[y_r, y_r + h]$, and $(y_r + h, \infty)$. You should think of h as a *very* small quantity which tends to zero.

The main idea of our proof is the following: let's compute the probability that $r-1$ values lie in the first interval, one lies in the middle and $n-r$ lie in the last. We would like to say that is just

$$\text{Prob}(Y_r \in [y_r, y_r + h]) \ = \ \int_{y_r}^{y_r+h} g_r(y_r) dy_r;$$

unfortunately, we can't. The problem is that we could have two or three or even more of the random variables in $[y_r, y_r + h]$; if for example exactly two were in the middle interval then our calculation above would be slightly off. This isn't too big of an issue, though. We'll see that it's very unlikely to have two or more values in this interval, as it's of size h with $h \to 0$. A careful analysis will show that we don't have to worry about this case in the limit. We'll then use the Mean Value Theorem to conclude that the integral is essentially $g_r(y_r)h$ plus something very very small, and from this we'll get the claimed density.

Now, to the details! We first find the probability that Y_r falls in the middle of these three intervals, and *no other value from the sample falls in this interval*. In order for this to be the case, we must have $r-1$ values falling in the first interval, one value falling in the second, and $n-r$ falling in the last interval. We need a basic fact about multinomial distributions (see **ADD REF??**). You should review that section if the argument here seems incomplete. We have n values, $X_1, \dots, X_n$. We need to choose exactly $r-1$ to lie in the first interval, 1 in the middle, and $n-r$ in the last. There are $\binom{n}{r-1}$ ways to choose the $r-1$ values to lie in the first interval. Of the remaining $n-(r-1)$ values, we have to choose 1 to lie in the middle; there are $\binom{n-(r-1)}{1}$ ways to do this. Finally, we have $n-r$ remaining values, all of which lie in the last interval; there is just $\binom{n-r}{n-r}$ ways to do that. Combining all these pieces, we have the combinatorial piece of the density: the number of ways to have exactly $r-1$ in the first interval, 1 in the middle and $n-r$ in the last is just

$$\begin{aligned}\binom{n}{r-1}\binom{n-(r-1)}{1}\binom{n-r}{n-r} \quad &= \quad \frac{n!}{(r-1)!(n-(r-1))!} \cdot \frac{(n-(r-1))!}{1!(n-r)!} \\ &= \quad \frac{n!}{(r-1)!(n-r)!};\end{aligned}$$

this is also the multinomial coefficient $\binom{n}{r-1,1,n-r}$ (though we don't really need this for our proof).

We've made good progress. We've shown that

$$\begin{aligned}&\mathrm{Prob}(Y_r \in [y_r, y_r + h] \text{ and } Y_i \neq [y_r, y_r + h] \text{ if } i \neq r) \ = \\ &\frac{n!}{(r-1)!1!(n-r)!}\left[\int_{-\infty}^{y_r} f(x)\,dx\right]^{r-1}\left[\int_{y_r}^{y_r+h} f(x)\,dx\right]^{1}\left[\int_{y_r+h}^{\infty} f(x)\,dx\right]^{n-r}.\end{aligned} \tag{1.1}$$

We need also consider the case of two or more of the Y_i lying in $[y_r, y_r + h]$. This is the most technical part of the argument. We'll briefly describe the answer now, and save the full details for later. The idea is that this probability must be small. What do we mean by small? We mean it must be of order h^2 or smaller, which we write as $O(h^2)$. This means there is some universal constant C such that, for all h sufficiently small, the probability is at most Ch^2. The probability any one value lies in this interval is just $\int_{y_r}^{y_r+h} f(x)dx$, and this integral is of size h; thus 'roughly' the probability of at least two being in this interval should be the square of this. Again, this is merely a sketch; we'll give the full details for those who wish them at the end. Let's just accept this for now and move on with the rest of the proof.

The consequence of the arguments above is that we may remove the constraint that exactly one $Y_i \in [y_r, y_r + h]$ in (1.1) at a cost of at most $O(h^2)$. For our purposes, it doesn't matter if there's exactly one or instead several values in this interval; as long as there is at least one then Y_r is in this interval. We thus find that

$$\begin{aligned}&\mathrm{Prob}(Y_r \in [y_r, y_r + h]) \ = \\ &\frac{n!}{(r-1)!1!(n-r)!}\left[\int_{-\infty}^{y_r} f(x)\,dx\right]^{r-1}\left[\int_{y_r}^{y_r+h} f(x)\,dx\right]^{1}\left[\int_{y_r+h}^{\infty} f(x)\,dx\right]^{n-r} \\ &\ + O(h^2).\end{aligned}$$

We now apply the Mean Value Theorem. If F is an anti-derivative of f, then the Mean Value Theorem applied to F can be written either as

$$\frac{F(b) - F(a)}{b - a} \ = \ F'(c)$$

or

$$\int_a^b f(x)dx \ = \ (b - a) \cdot f(c).$$

. We find that for some c_{h,y_r} with $y_r \leq c_{h,y_r} \leq y_r + h$, we have

$$\int_{y_r}^{y_r+h} f(x)\,dx \ = \ h \cdot f(c_{h,y_r}). \tag{1.2}$$

We denote the point provided by the Mean Value Theorem by c_{h,y_r} in order to emphasize its dependence on h and y_r.

We can substitute this result into the expression in (1.2). We divide the result by h (the length of the middle interval $[y_r, y_r + h]$), and consider the limit as $h \to 0$:

$$\begin{aligned}
&\lim_{h\to 0} \frac{\operatorname{Prob}(Y_r \in [y_r, y_r+h])}{h} \\
=\ &\lim_{h\to 0} \frac{\frac{n!}{(r-1)!1!(n-r)!}\left[\int_{-\infty}^{y_r} f(x)\,dx\right]^{r-1}\left[\int_{y_r+h}^{\infty} f(x)\,dx\right]^{1}\left[\int_{y_r+h}^{\infty} f(x)\,dx\right]^{n-r}}{h} \\
&\quad + \lim_{h\to 0}\frac{O(h^2)}{h} \\
=\ &\lim_{h\to 0} \frac{\frac{n!}{(r-1)!1!(n-r)!}\left[\int_{-\infty}^{y_r} f(x)\,dx\right]^{r-1} h\cdot f(c_{h,y_r})\left[\int_{y_r+h}^{\infty} f(x)\,dx\right]^{n-r}}{h} \\
=\ &\lim_{h\to 0} \frac{n!}{(r-1)!1!(n-r)!}\left[\int_{-\infty}^{y_r} f(x)\,dx\right]^{r-1} f(c_{h,y_r})\left[\int_{y_r+h}^{\infty} f(x)\,dx\right]^{n-r} \\
=\ &\frac{n!}{(r-1)!1!(n-r)!}\left[\int_{-\infty}^{y_r} f(x)\,dx\right]^{r-1} f(y_r)\left[\int_{y_r}^{\infty} f(x)\,dx\right]^{n-r}. \qquad (1.3)
\end{aligned}$$

Thus the proof is reduced to showing that the left hand side above is $g_r(y_r)$. Let $g_r(y_r)$ be the probability density of Y_r. Let $G_r(y_r)$ be the cumulative distribution function of Y_r. Then

$$\operatorname{Prob}(Y_r \le y) \ =\ \int_{-\infty}^{y} g_r(y_r)dy_r \ =\ G_r(y),$$

and $G_r'(y) = g_r(y)$. Thus the left hand side of (1.3) equals

$$\lim_{h\to 0}\frac{\operatorname{Prob}(Y_r \in [y_r, y_r+h])}{h} \quad=\quad \lim_{h\to 0}\frac{G_r(y_r+h) - G_r(y_r)}{h} \ =\ g_r(y_r),$$

where the last equality follows from the definition of the derivative. This completes the proof. □

The technique employed in this proof is a common method for calculating probability densities. We first calculate the probability that a random variable Y lies in an infinitesimal interval $[y, y+h]$. This probability is $G(y+h) - G(y)$, where g is the density of Y and G is the cumulative distribution function (so $G' = g$). The definition of the derivative yields

$$\lim_{h\to 0}\frac{\operatorname{Prob}(Y \in [y, y+h])}{h} \ =\ \lim_{h\to 0}\frac{G(y+h) - G(y)}{h} \ =\ g(y).$$

As promised, we now turn to showing that there is negligibly small probability of two or more values of the X_i's lying in the interval $[y_r, y_r + h]$. This part may be safely skipped, as it's the technical justification for one of the steps in the proof. Using the law of total probability, this is just 1 minus the probability that either none of the X_i's lie in this interval or exactly one does. Thus this probability is

$$1 - \left(\int_{x \notin [y_r, y_r+h]} f(x)dx\right)^n - \binom{n}{1}\left(\int_{x \in [y_r, y_r+h]} f(x)dx\right) \cdot \left(\int_{x \notin [y_r, y_r+h]} f(x)dx\right)^{n-1}. \tag{1.4}$$

To evaluate the above, we use

$$\int_{x \notin [y_r, y_r+h]} f(x)dx \;=\; 1 - \int_{x=y_r}^{y_r+h} f(x)dx.$$

It is here that we finally start using our assumptions. As we're assuming f is continuous, a result from analysis gives that it's bounded on the interval $[y_r, y_r+1]$ by some number, say B. The integral is therefore at most Bh; let's let $\mathcal{I}(h)$ denote the value of this integral:

$$\mathcal{I}(h) \;=\; \int_{x=y_r}^{y_r+h} f(x)dx.$$

Substituting into (1.4) gives that the probability of two or more of the X_i's being in the interval is

$$\begin{aligned} & 1-(1-\mathcal{I}(h))^n - n\mathcal{I}(h)\,(1-\mathcal{I}(h))^{n-1} \\ = \;& 1-\left(1-n\mathcal{I}(h)+\binom{n}{2}\mathcal{I}(h)^2+\cdots\right) - n\mathcal{I}(h)\left(1-\binom{n-1}{1}\mathcal{I}(h)+\cdots\right) \\ = \;& 1-1+n\mathcal{I}(h)-\binom{n}{2}\mathcal{I}(h)^2+\cdots-n\mathcal{I}(h)+n\binom{n-1}{1}\mathcal{I}(h)^2+\cdots \\ = \;& B_2\mathcal{I}(h)^2+B_3\mathcal{I}(h)^3+\cdots+B_n\mathcal{I}(h)^n. \end{aligned}$$

Notice that the constant term canceled, the $\mathcal{I}(h)$ term canceled, and all that remains is terms with $\mathcal{I}(h)$ to the second or higher power. We'll spare you the rest of the details, but clearly as $h \to 0$ the sum is dominated by the $\mathcal{I}(h)^2$ piece, and thus there is some constant B (which of course can depend on n, r, and y_r) such that the probability of two or more of the X_i's lying in $[y_r, y_r+h]$ is $O(h^2)$ as h tends to zero.

1.3 Examples of Order Statistics

After developing the general theory, it's fun to apply it to some special cases. Let's calculate the distribution of the largest and smallest values of n independent measurements from the uniform distribution on $[0,1]$. So $X_1, \dots, X_n$ are independent, identically distributed uniform random variables on $[0,1]$. Their density f satisfies

$$f(x) \;=\; \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

From Theorem 1.2.1, we find

$$g_1(y_1) \;=\; nf(y_1)\left(\int_{y_1}^{\infty} f(x)dx\right)^{n-1}.$$

Clearly $g_1(y_1) = 0$ unless $0 \leq y_1 \leq 1$; if that happens, then

$$g_1(y_1) \;=\; nf(y_1)(1-y_1)^{n-1} \;=\; n(1-y_1)^{n-1};$$

note the smaller y_1 is the larger the density (as one would expect). For Y_n, we argue similarly and find

$$g_n(y_n) \;=\; nf(y_n)\left(\int_{-\infty}^{y_n} f(x)dx\right)^{n-1} \;=\; nf(y_n)(y_n-0)^n \;=\; ny_n^n.$$

Not surprisingly, Y_r exhibits the opposite behavior to Y_1.

As always, there is a simple trick we can do to see if our answers above are consistent. Note our random variables are symmetric functions about $1/2$; in other words, the density satisfies $f(1/2+u) = f(1/2-u)$. One consequence of this is that X_i and $1 - X_i$ have the same distribution, and $\min(X_1, \dots, X_n) = \max(1 - X_1, \dots, 1 - X_n)$. To put it another way, the probability Y_1 equals a should be the same as the probability that Y_n equals $1-a$. In other words, we should have $g_1(1-y_n) = g_n(y_n)$; after doing a little algebra (including a change of variables) we see that we *do* have this, and happily all is consistent:

$$\begin{aligned}
g_1(1-y_n) &= nf(1-y_n)\left(\int_{1-y_n}^{\infty} f(x)dx\right)^{n-1} \\
&= nf(y_n)\left(\int_{y_n}^{-\infty} f(1-v)(-dv)\right)^{n-1} \\
&= nf(y_n)\left(\int_{-\infty}^{y_n} f(v)dv\right)^{n-1} \\
&= g_n(y_n).
\end{aligned}$$

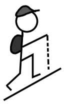
It's interesting to compute some properties of Y_1 (or, equivalently, Y_n). The most important, of course, is the mean or expected value. It is

$$\begin{aligned}
\mathbb{E}[Y_1] &= \int_0^1 y_1 g_1(y_1)dy_1 \\
&= \int_0^1 y_1 \cdot n(1-y_1)^{n-1}dy_1.
\end{aligned}$$

There are many ways to evaluate this integral. Probably the easiest is to integrate by parts. Let $u = y_1$, $du = dy_1$, $dv = n(1-y_1)^{n-1}dy_1$ and $v = -(1-y_1)^n$. Then

$$\begin{aligned}
\mathbb{E}[Y_1] &= uv\Big|_0^1 - \int_0^1 vdu \\
&= y_1(1-y_1)^n\Big|_0^1 + \int_0^1 (1-y_1)^n dy_1 \\
&= (0-0) + \frac{-(1-y_1)}{n+1}\Big|_0^1 \\
&= \frac{1}{n+1}.
\end{aligned}$$

Is this answer reasonable? The simplest case is if $n = 1$, in which case $Y_1 = X_1$. As X_1 is uniform on $[0, 1]$, its average value is 1/2, which is precisely what we get here. As n increases the mean decreases, which again is reasonable as Y_1 is the minimum of a larger and larger set, and thus it has a greater and greater chance of being small. As an aside, it's worth remarking that the expected value of Y_n is $\frac{n}{n+1}$.



For a nice exercise, see if you can find a nice formula for the minimum and the maximum of 2, 3 or 4 independent standard exponentials (i.e., $f(x) = e^{-x}$ for $x \geq 0$ and 0 otherwise). Can you generalize your answer?

1.4 The Sample Distribution of the Median

Okay. We've spent a lot of time and written down a multitude of explicit equations for all the order statistics, and we've done some calculations. It's time to get some dividends! Specifically, it's time to really see why this is such an important topic.

In addition to the smallest (Y_1) and largest (Y_n) order statistics, we're often interested in the **sample median**, $\tilde{X}$. For a sample of odd size, $n = 2m + 1$, the sample median is defined as Y_{m+1}. If $n = 2m$ is even, the sample median is defined as $\frac{1}{2}(Y_m + Y_{m+1})$.

As always, let's assume our random variable has a continuous density. We'll prove a relation between the sample median and the **population median** $\tilde{\mu}$. By definition, $\tilde{\mu}$ satisfies

$$\int_{-\infty}^{\tilde{\mu}} f(x)\,dx \;=\; \frac{1}{2}. \tag{1.5}$$

It is convenient to re-write the above in terms of the cumulative distribution function. If F is the cumulative distribution function of f, then $F' = f$ and (1.5) becomes

$$F(\tilde{\mu}) \;=\; \frac{1}{2}.$$

We are now ready to consider the distribution of the sample median. This is the big theorem of the chapter; as discussed, it serves as a good substitute to the Central Limit Theorem in many cases.

Theorem 1.4.1 (Sample Median) *Let a sample of size $n = 2m + 1$ with n large be taken from an infinite population with a density function $f(x)$ (in other words, we have n independent, identically distributed random variables with density f) that is nonzero at the population median $\tilde{\mu}$ and continuously differentiable in a neighborhood of $\tilde{\mu}$. The sampling distribution of the median is approximately normal with mean $\tilde{\mu}$ and variance $\frac{1}{8f(\tilde{\mu})^2 m}$.*

While this is an extremely important theorem, and one of the key applications of order statistics, the proof is long. It's frequently omitted in a first course in probability; there is no harm in skipping it. Because most books don't offer a proof, it can be frustrating for those who want to understand what's really going on, as you have to track down a proof somewhere. So, partly for completeness and, as always, partly to emphasize the techniques and methods of the proofs, we provide a proof below. As the argument is long, you should skim it first to take in the main points, and then if you desire go back and try to follow it step by step.

For those going forward with the proof, here's a quick outline of what you'll see; the list of topics should show you why we think the proof merits inclusion! We'll use Stirling's formula to estimate the factorials, Taylor series to expand our functions about the median, and the definition of e^x to understand the limiting behavior. This last step involves some delicate analysis, specifically handling how powers converge to exponentials. It's very easy to make a mistake here, and we'll show a nice technique to attack problems like this. We'll end by doing some algebra to help us recognize that the limit converges to being normally distributed.

Proof: Let the median random variable $\tilde{X}$ have values $\tilde{x}$ and density $g(\tilde{x})$. The median is simply the $(m+1)^{\text{th}}$ order statistic, so its distribution is given by Theorem 1.2.1:

$$g(\tilde{x}) \ = \ \frac{(2m+1)!}{m!m!}\left[\int_{-\infty}^{\tilde{x}} f(\tilde{x})\,dx\right]^m f(\tilde{x})\left[\int_{\tilde{x}}^{\infty} f(x)\,dx\right]^m.$$

We first find an approximation for the constant factor in this equation. For this, we use Stirling's formula (see **ADD REF**), which tells us that $n! = n^n e^{-n}\sqrt{2\pi n}(1 + O(n^{-1}))$. We'll consider values sufficiently large so that the terms of order $1/n$ need not be considered. Rather then clutter the already long argument with more careful book-keeping, we'll argue at this level of informality; the interested reader can go through and add the details. Hence

$$\begin{aligned}\frac{(2m+1)!}{m!m!} &= \frac{(2m+1)(2m)!}{(m!)^2}\\ &\approx \frac{(2m+1)(2m)^{2m}e^{-2m}\sqrt{2\pi(2m)}}{(m^m e^{-m}\sqrt{2\pi m})^2}\\ &= \frac{(2m+1)4^m}{\sqrt{\pi m}}.\end{aligned}$$

As F is the cumulative distribution function, $F(\tilde{x}) = \int_{-\infty}^{\tilde{x}} f(x)\,dx$, which implies

$$g(\tilde{x}) \ \approx \ \frac{(2m+1)4^m}{\sqrt{\pi m}}\,[F(\tilde{x})]^m\, f(\tilde{x})\,[1 - F(\tilde{x})]^m\,.$$

It's generally a good idea to replace complicated functions with simpler ones. Thus, we'll use the Taylor series expansion of $F(\tilde{x})$ about $\tilde{\mu}$, which is just

$$F(\tilde{x}) \ = \ F(\tilde{\mu}) + F'(\tilde{\mu})(\tilde{x} - \tilde{\mu}) + O((\tilde{x} - \tilde{\mu})^2).$$

Because $\tilde{\mu}$ is the population median, $F(\tilde{\mu}) = 1/2$. Further, since F is the cumulative distribution function, $F' = f$ and we find

$$F(\tilde{x}) \ = \ \frac{1}{2} + f(\tilde{\mu})(\tilde{x} - \tilde{\mu}) + O((\tilde{x} - \tilde{\mu})^2). \tag{1.6}$$

This approximation is only useful if $\tilde{x} - \tilde{\mu}$ is small; in other words, we need $\lim_{m\to\infty} |\tilde{x} - \tilde{\mu}| = 0$. Fortunately this is easy to show, and a proof is included at the end of this section. So as not to lose the flow of the argument, let's assume this holds for now and continue.

Let $t = \tilde{x} - \tilde{\mu}$ (which is small and tends to 0 as $m \to \infty$). We want to substitute our Taylor series expansion into (1.6). Actually, we need to expand even further – this is sadly an instance where the standard, first order Taylor series does not provide enough control! The problem is each term has an error of size $O(t^2)$. When we multiply them together there is also an error of size $O(t^2)$, and this is the same order of magnitude as the secondary term, $(f(\widetilde{\mu})t)^2$. The remedy is to be more careful in expanding $F(\tilde{x})$ and $1 - F(\tilde{x})$. A careful analysis shows that their t^2 terms are equal in magnitude but opposite in sign. Thus they will cancel in the calculations below. In summary, we really need to use $F(\tilde{x}) = \frac{1}{2} + f(\tilde{\mu})(\tilde{x} - \tilde{\mu}) + \frac{f'(\tilde{\mu})}{2}(\tilde{x} - \tilde{\mu})^2$ plus an error of size $O((x - \widetilde{\mu})^3)$ (and similarly for $1 - F(\tilde{x})$). We find

$$\begin{aligned} g(\tilde{x}) \approx \frac{(2m+1)4^m}{\sqrt{\pi m}} & \left[\frac{1}{2} + f(\tilde{\mu})t + \frac{f'(\tilde{\mu})}{2}(\tilde{x} - \tilde{\mu})^2 + O(t^3)\right]^m \\ f(\tilde{x}) & \left[1 - \left(\frac{1}{2} + f(\tilde{\mu})t + \frac{f'(\tilde{\mu})}{2}(\tilde{x} - \tilde{\mu})^2 + O(t^3)\right)\right]^m . \end{aligned}$$

By rearranging and combining factors, we find that

$$\begin{aligned} g(\tilde{x}) &\approx \frac{(2m+1)4^m}{\sqrt{\pi m}} f(\tilde{x}) \left[\frac{1}{4} - (f(\tilde{\mu})t)^2 + O(t^3)\right]^m \\ &= \frac{(2m+1)f(\tilde{x})}{\sqrt{\pi m}} \left[1 - \frac{4m(f(\tilde{\mu})t)^2}{m} + O(t^3)\right]^m . \end{aligned} \tag{1.7}$$

Remember that one definition of e^x is

$$e^x = \exp(x) = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n ;$$

see **ADD REF ??** for a review of properties of the exponential function. It's not immediately clear that we can use this, as we have m in both the numerator and the denominator; however, t is supposed to be quite small, and we'll show later it's so small that mt^2 cannot be that large with respect to m. So, for now let's not worry about justifying this step and just use the exponential relation. While we're at it, let's also ignore higher powers of t for the moment. We find for large m that

$$\begin{aligned} g(\tilde{x}) &\approx \frac{(2m+1)f(\tilde{x})}{\sqrt{\pi m}} \exp\left(-4mf(\tilde{\mu})^2 t^2\right) \\ &\approx \frac{(2m+1)f(\tilde{x})}{\sqrt{\pi m}} \exp\left(-\frac{(\tilde{x} - \tilde{\mu})^2}{1/(4mf(\tilde{\mu})^2)}\right). \end{aligned} \tag{1.8}$$

Since, as shown below, $\tilde{x}$ can be assumed arbitrarily close to $\tilde{\mu}$ with high probability, we can assume $f(\tilde{x}) \approx f(\tilde{\mu})$. To prove that there is negligible error in replacing $f(\tilde{x})$ with $f(\tilde{\mu})$, we use the mean value theorem and find

$$f(\tilde{x}) - f(\tilde{\mu}) = f'(c_{\tilde{x},\tilde{\mu}}) \cdot (\tilde{x} - \tilde{\mu});$$

here we have written the constant as $c_{\tilde{x},\tilde{\mu}}$ to emphasize the fact that we evaluate the first derivative in the interval $[\tilde{x}, \tilde{\mu}]$. As we have assumed f is continuously differentiable and $|\tilde{x} - \tilde{\mu}|$ is small, we may bound $f'(c_{\tilde{x},\tilde{\mu}})$ Thus we may replace

$f(\tilde{x})$ with $f(\tilde{\mu})$ at a cost of $O(t)$, where $t = \tilde{x} - \tilde{\mu}$ tends to zero with m. We therefore have

$$g(\tilde{x}) \approx \frac{(2m+1)f(\tilde{\mu})}{\sqrt{\pi m}} \exp\left(-\frac{(\tilde{x}-\tilde{\mu})^2}{1/(4mf(\tilde{\mu})^2)}\right).$$

Looking at the exponential part of the expression for $g(\tilde{x})$, we see that it appears to be a normal density with mean $\tilde{\mu}$ and $\sigma^2 = 1/(8mf(\tilde{\mu})^2)$. If we were instead to compute the variance from the normalization constant, we would find the variance to be

$$\frac{m}{2(2m+1)^2 f(\tilde{\mu})^2}$$

We see that the two values are asymptotically equivalent, thus we can take the variance to be $\sigma^2 = 1/(8mf(\tilde{\mu})^2)$. Thus to complete the proof of the theorem, all that we need to is prove that we may ignore the higher powers of t and replace the product with an exponential in passing from (1.7) to (1.8). Using the identity

$$w^m = \exp\left(\log(w^m)\right) = \exp\left(m \log w\right).$$

We have

$$\left(1 - \frac{4m(f(\tilde{\mu})t)^2}{m} + O(t^3)\right)^m = \exp\left(m \log\left(1 - 4(f(\tilde{\mu})t)^2 + O(t^3)\right)\right). \tag{1.9}$$

We use the Taylor series expansion of $\log(1-x)$:

$$\log(1-x) = -x + O(x^2);$$

we only need one term in the expansion as t is small. Thus (1.9) becomes

$$\begin{aligned} \left(1 - \frac{4m(f(\tilde{\mu})t)^2}{m} + O(t^3)\right)^m &= \exp\left(-m \cdot 4(f(\tilde{\mu})t)^2 + O(mt^3)\right) \\ &= \exp\left(-\frac{(\tilde{x}-\tilde{\mu})^2}{1/(4mf(\tilde{\mu})^2)}\right) \cdot \exp(O(mt^3)). \end{aligned}$$

In §1.5 we'll show that as $m \to \infty$, $mt^3 \to 0$. Thus the $\exp(O(mt^3))$ term above tends to 1, which completes the proof. □

Our justification of ignoring the higher powers of t and replacing the product with an exponential in passing from (1.7) to (1.8) is a standard technique. As it's so important, we repeat again why we can do this. We are just replacing some quantity $(1-P)^m$ with $(1-P)^m = \exp(m\log(1-P))$. We then Taylor expand the logarithm, and look at the limit as $m \to \infty$.

ADD REF TO CLT CHAPTER WHERE CAN GO WRONG – PRODUCT ONE TO INFINITY ONE TO ZERO

1.5 Technical bounds for proof of Median Theorem

In proving the Median Theorem, we assumed that we could ignore higher powers of $t = \tilde{X} - \tilde{\mu}$. We are able to do this because, with high probability, t is small. Here we provide a more formal statement of this fact, as well as a proof.

Lemma 1.5.1 *Suppose $f(x)$ is a continuously differentiable function in some neighborhood of $\tilde{\mu}$. Then for any $c > 0$, we have*

$$\lim_{m\to 0} \mathrm{Prob}(|\tilde{X} - \tilde{\mu}| \geq c) = 0.$$

Proof: This is equivalent to proving that

$$\lim_{m\to 0} \mathrm{Prob}(\tilde{X} \leq \tilde{\mu} - c) = 0 \quad \text{and} \quad \lim_{m\to 0} \mathrm{Prob}(\tilde{X} \geq \tilde{\mu} + c) = 0.$$

We will prove only the first of these two statements as the proof of the second is very similar.

By (1.6), we can approximate the density of the median as

$$g(\tilde{x}) \approx \frac{(2m+1)4^m f(\tilde{x})}{\sqrt{\pi m}} \left([F(\tilde{x})]\,[1 - F(\tilde{x})]\right)^m.$$

We consider the factor $\left([F(\tilde{x})]\,[1 - F(\tilde{x})]\right)^m$. It is convenient to write $\theta = F(\tilde{x})$ and consider the function $h(\theta) = \theta(1-\theta)$. This function will attain its maximum for the same value of $\theta = F(\tilde{x})$ as $\left([F(\tilde{x})]\,[1 - F(\tilde{x})]\right)^m$, and it's a simple exercise in calculus to show that this value is $\theta = 1/2$. This condition holds only for $\tilde{x} = \tilde{\mu}$. We furthermore note that for $\theta < 1/2$, $h'(\theta) = 1 - 2\theta > 0$, so h is increasing. Since $F(\tilde{x}) = \frac{1}{2}$ precisely when $\tilde{x} = \tilde{\mu}$, this means that for $\tilde{x} \leq \tilde{\mu} - c$, the maximum value of $g(\theta)$ occurs for $\tilde{x} = \tilde{\mu} - c$. We therefore have for $\tilde{x} \leq \tilde{\mu} - c$,

$$\begin{aligned}
\left(F(\tilde{x})\,[1 - F(\tilde{x})]\right)^m &\leq \left(F(\tilde{\mu} - c)\,[1 - F(\tilde{\mu} - c)]\right)^m \\
&< \left(F(\tilde{\mu})\,[1 - F(\tilde{\mu})]\right)^m = \frac{1}{4^m}.
\end{aligned} \tag{1.10}$$

We choose α so that $\frac{\alpha}{4} = F(\tilde{\mu} - c)\,(1 - F(\tilde{\mu} - c))$. Equation (1.10) then tells us that for $\tilde{x} \leq \tilde{\mu} - c$,

$$\left(F(\tilde{x})\,[1 - F(\tilde{x})]\right)^m \leq \left(\frac{\alpha}{4}\right)^m \leq \frac{1}{4^m}.$$

In particular, we note that $\alpha < 1$.

We now begin to look at the probability that $\tilde{X}$ is at most $\tilde{\mu} - c$. We have

$$\begin{aligned}
\mathrm{Prob}(\tilde{X} \leq \tilde{\mu} - c) &= \int_{-\infty}^{\tilde{\mu}-c} g(\tilde{x})\, d\tilde{x} \\
&\approx \int_{-\infty}^{\tilde{\mu}-c} \frac{(2m+1)4^m}{\sqrt{\pi m}} f(\tilde{x}) F(\tilde{x})^m (1 - F(\tilde{x}))^m\, d\tilde{x} \\
&< \frac{(2m)4^m}{\sqrt{m}} \int_{-\infty}^{\tilde{\mu}-c} f(\tilde{x}) F(\tilde{x})^m (1 - F(\tilde{x}))^m\, d\tilde{x}.
\end{aligned}$$

In the last step, we use the fact that for m sufficiently large ($m > 1$, in fact), $2m < \frac{2m+1}{\sqrt{\pi}}$. This simplifies the expression as a factor of $2m$ is easier to work with than the factor of $2m+1$. We now apply our bound on $F(\tilde{x})(1-F(\tilde{x}))$ to find that

$$\begin{aligned}\mathrm{Prob}(\tilde{X} \leq \tilde{\mu} - c) &< \frac{(2m)4^m}{\sqrt{m}} \int_{-\infty}^{\tilde{\mu}-c} f(\tilde{x}) \left(\frac{\alpha}{4}\right)^m d\tilde{x} \\ &= \frac{(2m)4^m}{\sqrt{m}} \left(\frac{\alpha}{4}\right)^m \int_{-\infty}^{\tilde{\mu}-c} f(\tilde{x})\, d\tilde{x} \\ &< 2\alpha^m \sqrt{m} \int_{-\infty}^{\tilde{\mu}} f(\tilde{x})\, d\tilde{x}.\end{aligned}$$

In obtaining the rightmost expression, we have used the fact that $f(\tilde{x})$ is nonnegative everywhere and positive in a neighborhood of $\tilde{\mu}$, so that $\int_{-\infty}^{\tilde{\mu}-c} f(\tilde{x})\, d\tilde{x} < \int_{-\infty}^{\tilde{\mu}} f(\tilde{x})\, d\tilde{x}$. Since $\tilde{\mu}$ is the median of the population, by definition, we have $\int_{-\infty}^{\tilde{\mu}} f(\tilde{x})\, d\tilde{x} = \frac{1}{2}$, so that

$$\mathrm{Prob}(\tilde{X} \leq \tilde{\mu} - c) \;<\; \alpha^m \sqrt{m}.$$

Since $\alpha < 1$, it follows that the right side of this inequality must converge to 0 as m goes to infinity, so the probability on the right side must likewise converge to 0. $\square$

Problem 1.5.2 *Let* $\alpha \in (0,1)$. *Prove*

$$\lim_{m\to\infty} \alpha^m \sqrt{m} \;=\; 0.$$

In fact, this expression tends to zero exponentially fast. Let $\delta = 1 - \alpha$. *Show that for* m *sufficiently large,*

$$\alpha^m \sqrt{m} \;\leq\; A\left(1 - \frac{\delta}{2}\right)^m \;=\; Ae^{-Bm},$$

where A *and* B *are constants (with* $B = \log\left(1 - \frac{\delta}{2}\right)$.

1.6 The Median of Normal Random Variables

Consider the case of a normal population with mean μ and variance σ^2. The normal density is symmetric about the mean $\tilde{\mu}$, hence $\tilde{\mu} = \mu$. Furthermore, we have

$$\begin{aligned}f(\tilde{\mu}) &= f(\mu) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mu-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}},\end{aligned}$$

which implies that

$$\frac{1}{8mf(\tilde{\mu})^2} \;=\; \frac{\pi\sigma^2}{4m}.$$

For large $n = 2m + 1$, we therefore see that the distribution of the median (from a normal distribution with mean μ and variance σ^2) will be approximately normal with mean μ and variance $\pi\sigma^2/4m$.

What can we say about the population mean? Either from the Central Limit Theorem, or calculating directly as the sum of normal random variables is normal, we see that the sample mean is normally distributed with mean μ and variance $\sigma^2/n = \sigma^2/(2m+1)$.

So, both the sample median and the sample mean converge to being normally distributed with mean μ; however, the sample median has variance approximately $\pi\sigma^2/4m$, while the mean has variance $\sigma^2/(2m+1)$. If we take the ratio of the two, we find

$$\frac{\text{Variance of Sample Median}}{\text{Variance of Sample Mean}} \approx \frac{\pi\sigma^2/4m}{\sigma^2/(2m+1)} = \frac{\pi}{2}\frac{2m+1}{2m}.$$

For large m, this ratio is about $\pi/2 \approx 1.57$. What does this mean? It means that while the sample median and sample mean have the same expected value, the sample median has larger fluctuations. How should we interpret this? The cleanest is that the sample mean will do a better job; it's 'tighter' about the true value.

The moral of this example is that, as nice as the Median Theorem is, it's most emphatically *not* a replacement for the Central Limit Theorem whenever the mean and the median are equal. The Central Limit Theorem can give better results; the Median Theorem should be kept in reserve for situations where the Central Limit Theorem isn't applicable (for instance, when we have a distribution like the Cauchy which has infinite variance, and possibly even undefined mean!).