MTH 224: Probability and Statistics **Fall 2017**
Course Notes 3 **Drew Armstrong**

Nov 9

We have finished the first two thirds of the course, which in retrospect I would label as

Part I: Introduction to Probability

Part II: Random Variables

I anticipate that the final third of the course will fit under the label

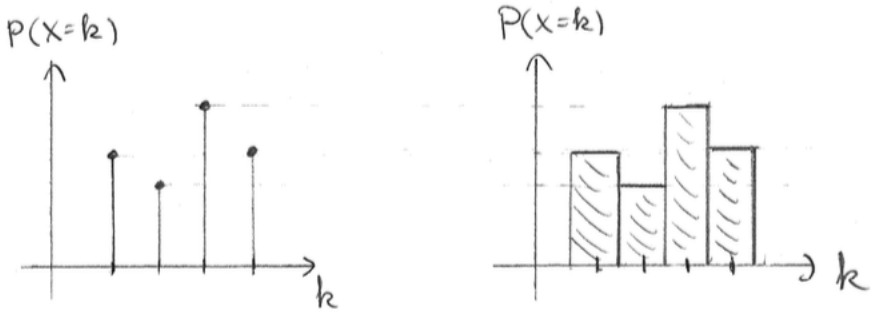
Part III: Introduction to Statistics

I have hinted at statistical ideas from time to time but we didn't address them fully because we didn't have the mathematical tools. We are almost ready to address real statistical problems but we need one final mathematical tool: the idea of a *continuous random variable*.

Recall that a discrete random variable X has a *probability mass function* (pmf) defined by

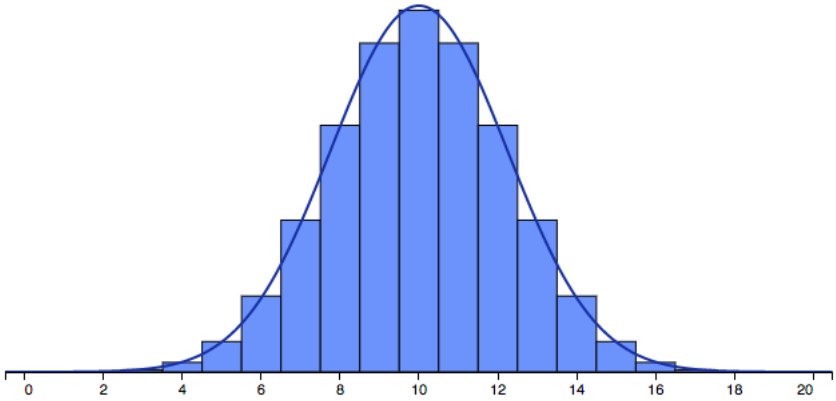
$$f_X(k) = P(X = k).$$

We like to draw pictures of pmfs as *line graphs* or *histograms*:



In the line graph we obtain the probability of an event by summing the **lengths** of the corresponding line segments. In the histogram we obtain the probability of an event by summing the **areas** of the corresponding rectangles.

For certain histograms, such as the binomial, we observe that the bars of the histogram start to resemble the area under a smooth curve. For example, here is the histogram for a binomial random variable X with $n = 20$ and $p = 1/2$:



As shown in the picture, it seems that the bars of this histogram can be approximated by the area under a certain "bell-shaped" curve. At the moment we don't know any details about this curve. Let's just assume that it's the graph of a certain real-valued function $g : \mathbb{R} \to \mathbb{R}$.

Now let's brush off our Calculus skills. If we want the exact probability that X is between 9 and 12 (inclusive) then we need to add the areas of the corresponding rectangles:

$$P(9 \leqslant X \leqslant 12) = P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12).$$

But if are happy with an approximate value then we might replace these rectangles by the area under the smooth curve between 8.5 and 12.5:

$$P(9 \leqslant X \leqslant 12) \approx \int_{8.5}^{12.5} g(k)\, dk.$$

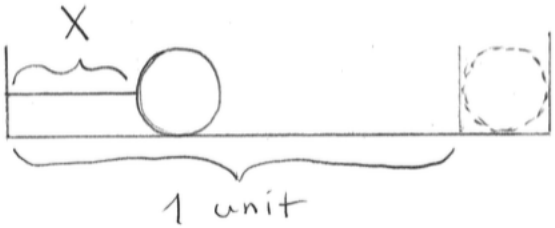
The reason we use the endpoints 8.5 and 12.5 instead of 9 and 12 is because the rectangle centered on $k = 9$ has its left endpoint at $k = 8.5$ and the rectangle centered on $k = 12$ has its right endpoint at $k = 12.5$. We still get an approximation if we integrate from 9 to 12, but it won't be as accurate.

We will discuss the details of this bell-shaped curve in the next lecture. For today, let me prepare the way by introducing the general concept of a continuous random variable.

Continuous random variables behave very much like discrete random variables except for one crucial difference: **a continuous random variable does not have a probability mass function.** In fact, if $X : S \to \mathbb{R}$ is a continuous random variable then we will find that the probability of any single value is zero:

$$P(X = k) = 0 \quad \text{for all possible values of } k.$$

Example. Suppose we throw a ball at random onto a billiard table and wait for it to settle. Then we let X be its distance from a fixed wall. For convenience, let us assume that the largest possible value of X is 1 unit:



As before we define the *support of* X as the set $S_X \subseteq \mathbb{R}$ of all possible values that X can take. In this case the support is the closed interval from 0 to 1:

$$S_X = [0,1] = \{k \in \mathbb{R} : 0 \leqslant k \leqslant 1\}.$$

Now here's an interesting question.

Question: What is the probability that X is **exactly** equal to $1/2$? Is it zero or nonzero?

$$P(X = 1/2) = 0 \qquad \text{or} \qquad P(X = 1/2) \neq 0\,?$$

The answer to this question depends on the accuracy of our measuring equipment. If we are measuring X with a ruler then we might have to round to the nearest tick mark. In that case $P(X = 1/2)$ will be a nonzero number. However, if we suppose that our measuring equipment is arbitrarily accurate then I claim that $P(X = 1/2) = 0$.

To examine this claim, let us assume that $P(X = 1/2)$ takes some fixed non-negative value:

$$P(X = 1/2) = \varepsilon \geqslant 0.$$

There is nothing so special about the point $X = 1/2$ so we might as well assume that any point on the table is equally likely. In particular, we will assume that

$$P(X = 1/4) = \varepsilon, \quad P(X = 1/8) = \varepsilon, \quad P(X = 1/16) = \varepsilon, \quad \text{etc.}$$

But Kolmogorov's three rules of probability must still hold even in the continuous case. Since the points $X = 1/2$, $X = 1/4$, $X = 1/8$ are a subset of the full interval, Kolmogorov's rules tell us that

$$\begin{aligned} \{1/2, 1/4, 1/8, \ldots\} &\subseteq [0,1], \\ P\left(X \in \{1/2, 1/4, 1/8, \ldots\}\right) &\leqslant P(X \in [0,1]) \\ P(X = 1/2) + P(X = 1/4) + P(X = 1/8) + \cdots &\leqslant 1 \\ \varepsilon + \varepsilon + \varepsilon + \cdots &\leqslant 1. \end{aligned}$$

In other words, if we add the number ε to itself an infinite number of times then we obtain a number less than or equal to 1. This is clearly impossible unless $\varepsilon = 0$.

Thus, in order to preserve the rules of probability we must have $P(X = k) = 0$ for any specific value of k. ///

And yet, we also believe that the probabiliy of X falling in a given interval should equal the **length** of that interval:

$$\begin{aligned} P(0 \leqslant X \leqslant 1) &= 1, \\ P(0 \leqslant X \leqslant 1/2) &= 1/2, \\ P(0 \leqslant X \leqslant 1/3) &= 1/3, \\ P(1/3 \leqslant X \leqslant 2/3) &= 1/3. \end{aligned}$$

So how does this work?

Definition of Probability Density Function. Our fundamental analogy is that

$$\textit{probability} \approx \textit{mass}.$$

For discrete random variables we view probabiliy as a finite or infinite sum of point masses:

$$P(X \in A) = \sum_{k \in A} P(X = k).$$

However, if X is a continuous random variable then this definition does not work because we must have $P(X = k) = 0$ for any fixed k.[1] Instead, we will define a continuous random variable in terms of its density. To be specific, a continuous random variable $X : S \to \mathbb{R}$ is defined by a real-valued function

$$f_X : \mathbb{R} \to \mathbb{R}$$

which represents the **density** of X on the real line. Accordingly, we call this f_X the *probability density function (pdf)* of the random variable. To find the probability that X lies in a given interval $[a, b] \subseteq \mathbb{R}$ we **integrate the density** over this interval:

$$\begin{aligned} P(a \leqslant X \leqslant b) &= \int_a^b f_X(x)\, dx \\ \textit{mass} &= \int \textit{density}. \end{aligned}$$

In particular, for any fixed value k we find that

$$P(X = k) = P(k \leqslant X \leqslant k) = \int_k^k f_X(x)\, dx = 0,$$

which agrees with our previous discussion. ///

[1] You have to add up a **lot** of zeroes to get something that is nonzero!

In our billiard ball example, we take the density to be constantly equal to 1 inside the interval $[0, 1]$ and constantly equal to 0 outside this interval:

$$f_X(x) = \begin{cases} 1 & \text{for all } 0 \leqslant x \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

In particular, we find that

$$P(0 \leqslant X \leqslant 1) = \int_0^1 f_X(x)\,dx = \int_0^1 1\,dx = x\Big|_0^1 = 1 - 0 = 1.$$

And for any numbers $0 \leqslant k_1 \leqslant k_2 \leqslant 1$ we find that

$$P(k_1 \leqslant X \leqslant k_2) = \int_{k_1}^{k_2} f_X(x)\,dx = \int_0^1 1\,dx = x\Big|_{k_1}^{k_2} = k_2 - k_1.$$

This agrees with our previous intuition about this random variable.

Definition of Expected Value and Variance. Once we have made the transition from probability mass functions (pmfs) to probability density functions (pdfs), the rest of the theory of random variables goes through exactly as before.

If $f_X : \mathbb{R} \to \mathbb{R}$ is the pdf of a continuous random variable then we define its mean / expected value by the formula[2]

$$\mu_X = E[X] = \int x \cdot f_X(x)\,dx$$

and more generally we define the r-th moment by the formula

$$E[X^r] = \int x^r \cdot f_X(x)\,dx.$$

The expected value still represents the "center of mass" and the function $E[-]$ is still linear. The variance is defined as before:

$$\mathrm{Var}(X) = E[(X - \mu_X)^2] = \int (x - \mu_X)^2 \cdot f_X(x)\,dx.$$

And it can still be computed with the same trick:

$$\mathrm{Var}(X) = E[X^2] - E[X]^2.$$

///

Let's test out these formulas with an example.

[2]You should compare this to the formula for probability mass functions: $E[X] = \sum k \cdot P(X = k)$.

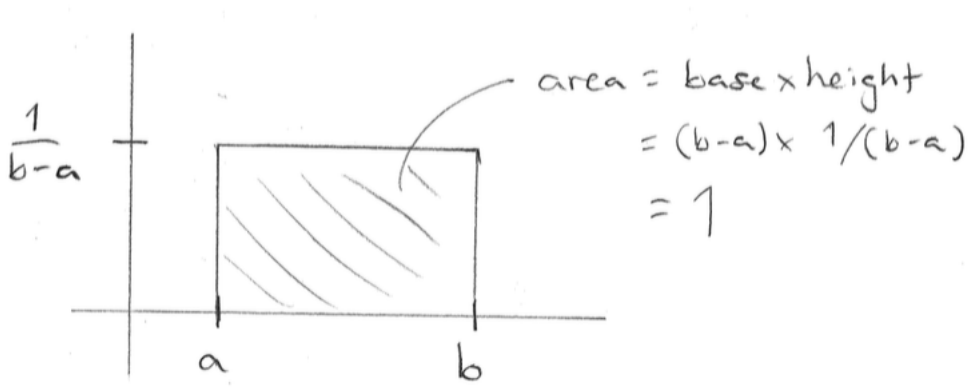
Example (Uniform Random Variables). Consider any real numbers $a < b$. We define the *uniform random variable* X on the interval $[a, b]$ by the function

$$f_X(x) = \begin{cases} 1/(b-a) & \text{for } a \leqslant x \leqslant b, \\ 0 & \text{otherwise.} \end{cases}$$

We observe that this is a probability density function because

$$\int_{-\infty}^{\infty} f_X(x)\,dx = \int_a^b \frac{1}{b-a}\,dx = \frac{x}{b-a}\bigg|_a^b = \frac{b}{b-a} - \frac{a}{b-a} = \frac{b-a}{b-a} = 1.$$

We don't even need Calculus to compute this integral because it is just the area of a rectangle:



Since this distribution is symmetric, I expect that the mean is the midway point between a and b, i.e., that $E[X] = (a+b)/2$. Let's verify this by computing the integral:

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx \\ &= \int_a^b x \cdot \frac{1}{b-a}\,dx \\ &= \frac{x^2}{2\cdot(b-a)}\bigg|_a^b \\ &= \frac{b^2-a^2}{2\cdot(b-a)} \\ &= \frac{\cancel{(b-a)}(b+a)}{2\cdot\cancel{(b-a)}} = \frac{a+b}{2}. \end{aligned}$$

That was a bit trickier but at least we knew what answer to expect. When computing the variance we **don't** know what to expect so we have to just trust the Calculus. First we compute the second moment:

$$E[X^2] = \int_{-\infty}^{\infty} x^2 \cdot f_X(x)\,dx$$

$$
\begin{aligned}
&= \int_a^b x^2 \cdot \frac{1}{b-a}\, dx \\
&= \frac{x^3}{3 \cdot (b-a)}\Big|_a^b \\
&= \frac{b^3 - a^3}{3 \cdot (b-a)} \\
&= \frac{\cancel{(b-a)}(a^2+ab+b^2)}{3 \cdot \cancel{(b-a)}} = \frac{a^2+ab+b^2}{3}.
\end{aligned}
$$

Here I used a general formula[3] for a difference of cubes:

$$\boxed{b^3 - a^3 = (b-a)(a^2+ab+b^2).}$$

Finally, we can compute the variance and the standard deviation. We have

$$
\begin{aligned}
\sigma_X^2 = \mathrm{Var}(X) &= E[X^2] - E[X]^2 \\
&= \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 \\
&= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} \\
&= \frac{4(a^2+ab+b^2)}{12} - \frac{3(a^2+2ab+b^2)}{12} \\
&= \frac{a^2-2ab+b^2}{12} = \frac{(a-b)^2}{12},
\end{aligned}
$$

and hence

$$\sigma_X = \sqrt{\frac{(a-b)^2}{12}} = \frac{b-a}{\sqrt{12}} \approx 0.289 \cdot (b-a).$$

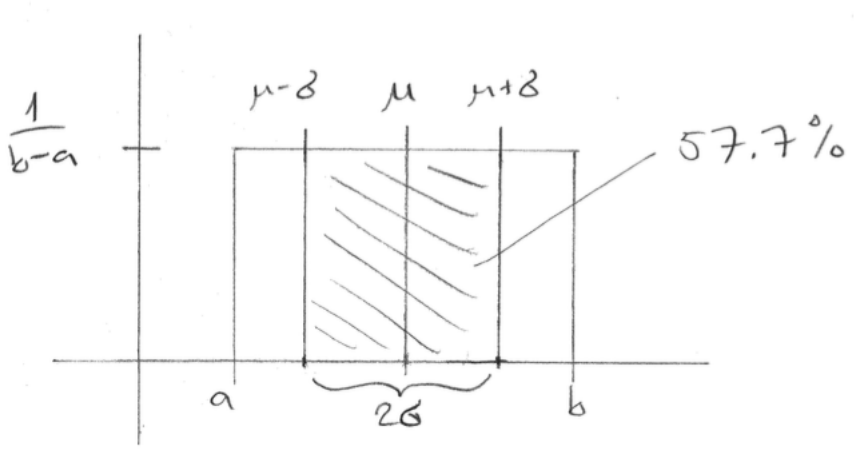
(Here I used the fact that $\sqrt{(a-b)^2} = (b-a)$ because $a < b$.) We conclude that the standard deviation is approximately 28.9% of the width of the interval.

Question: What is the probability that X is within one standard deviation of its mean?

Answer: $P(\mu_X - \sigma_X \leqslant X \leqslant \mu_X + \sigma_X) \approx 57.7\%$.

We could compute this by integrating the constant density $1/(b-a)$ from $x = \mu_X - \sigma_X$ to $x = \mu_X + \sigma_X$, but it is easier to think of this integral as the area of a rectangle:

[3] No, you do not need to memorize this formula. But you should multiply out the right hand side to verify that it is true.



Observe that the rectangle has height $1/(b-a)$ and width 2σ, so that

$$\text{(area)} = \text{(base)} \times \text{(height)} = 2\sigma \cdot \frac{1}{b-a} = \frac{2 \cdot \cancel{(b-a)}}{\sqrt{12}} \cdot \frac{1}{\cancel{b-a}} = \frac{2}{\sqrt{12}} \approx 57.7\%.$$

Nov 14

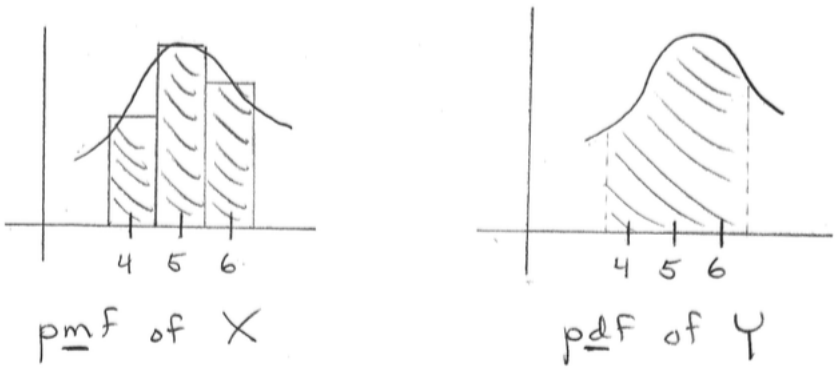
Last time I introduced the notion of a continuous random variable, which is defined in terms of its probability **density**[4] function. Here is a dictionary comparing various concepts for discrete and continuous random variables:

Discrete $X : S \to \mathbb{R}$	Continuous $X : S \to \mathbb{R}$
probability **mass** function $\sum_{k=-\infty}^{\infty} f_X(k) = 1$	probability **density** function $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$
probability of an **event** $A \subseteq \mathbb{R}$ $P(X \in A) = \sum_{k \in A} f_X(k)$	probability of an **interval** $[a,b] \subseteq \mathbb{R}$ $P(a \leqslant X \leqslant b) = \int_a^b f_X(x)\,dx$
probability of a single value $k \in \mathbb{R}$ $P(X = k) = f_X(k)$	probability of a single value $k \in \mathbb{R}$ $P(X = k) = P(k \leqslant X \leqslant k) = \int_k^k f_X(x)\,dx = 0$
center of mass / expected value $E[X] = \sum_{k=-\infty}^{\infty} k \cdot f_X(k)$	center of mass / expected value $E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx$

Except for these differences the discrete and continuous theories are completely parallel. The two theories can be compared directly if we think of a discrete random variable in terms of its histogram. For example, suppose that X, Y are random variables where X is **discrete** and

[4] As opposed to a probability **mass** function.

Y is **continuous**. Suppose further that we have $f_X(k) \approx f_Y(k)$ for all whole numbers $k \in \mathbb{R}$. Then the graph of f_Y is a close fit to the histogram of X, as in the following picture:



In this picture we also see that the probability of $X \in \{4, 5, 6\}$, i.e., the area of these three rectangles, is approximately equal to the area under the graph of f_Y from 3.5 to 6.5:

$$P(4 \leqslant X \leqslant 6) \approx P(3.5 \leqslant Y \leqslant 6.5)$$
$$\sum_{k=4}^{6} f_X(k) \approx \int_{3.5}^{6.5} f_Y(x)\,dx$$

Notice that we integrated from $3.5 = 4 - 0.5$ to $6.5 = 6 + 0.5$ instead of from 4 to 6 so that our region more closely matches the bars of the histogram. This trick is called the "continuity correction." It is not strictly necessary but it leads to better approximations.

Today we will discuss the first and most important example of this kind of discrete-continuous approximation, which was first discovered by Abraham de Moivre in the 1730s.

De Moivre's Problem. Suppose that a fair coin is flipped 3600 times and let X be the number of heads. What is the probability that X is between 1770 and 1830? ///

Since X is a binomial random variable with $n = 3600$ and $p = 1/2$ we can easily write down an exact formula for the probability:

$$\begin{aligned} P(1770 \leqslant X \leqslant 1830) &= \sum_{k=1770}^{1830} P(X = k) \\ &= \sum_{k=1770}^{1830} \binom{3600}{k} \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{3600-k} \end{aligned}$$

$$= \sum_{k=1770}^{1830} \binom{3600}{k} / 2^{3600}.$$

However, the numerators and denominators here are so **gigantic** that it is impossible to simplify this expression without a computer. In particular, the denominator has almost 2500 decimal digits:

$$\log_{10}\left(2^{3600}\right) \approx 2495.$$

De Moivre did not have access to a computer in 1730, but he did have a mastery of the relatively new techniques of Calculus.[5] By applying these techniques and some clever tricks he was able to evaluate the sum by hand to an accuracy of four decimal places:[6]

$$P(1770 \leqslant X \leqslant 1830) = \sum_{k=1770}^{1830} \binom{3600}{k} / 2^{3600} \approx 69.07\%.$$

That's pretty amazing! How did he do it?

First he made the change of variables $\ell = k - 1800$ so the sum is symmetric about zero:

$$\sum_{k=1770}^{1830} \binom{3600}{k} / 2^{3600} = \sum_{\ell=-30}^{30} \binom{3600}{1800+\ell} / 2^{3600}.$$

It is not a coincidence that 1800 is the expected value of X and 30 is the standard deviation.[7] In fact, de Moivre **invented** the concept of "standard deviation" through his study of this problem. The main problem here is that the binomial coefficient $\binom{3600}{1800+\ell}$ is only defined for whole numbers ℓ. To apply the techniques of Calculus, de Moivre needed to approximate this "discrete" expression with a "continuous" expression that is defined for all $\ell \in \mathbb{R}$.

I'll show you his solution in the more general case when a fair coin is flipped $2n$ times, with mean $\mu = n$ and standard deviation $\sigma = \sqrt{n/2}$. Afterwards we'll return to the specfic example where $2n = 3600$, $\mu = 1800$ and $\sigma = \sqrt{900} = 30$.

Step 1. *When the ratio ℓ/n is small*, de Moivre used some Calculus tricks to show that

$$\binom{2n}{n+\ell} / \binom{2n}{n} \approx e^{-\ell^2/n}.$$

Hence probability of getting $n + \ell$ heads has the following approximation:

$$P(X = n+\ell) = \binom{2n}{n+\ell} / 2^{2n} \approx e^{-\ell^2/n} \left[\binom{2n}{n} / 2^{2n} \right] = e^{-\ell^2/n} \cdot P(X = n).$$

Observe that the expression on the right makes sense for *any real number* $\ell \in \mathbb{R}$. ///

[5] Which was invented by Newton and Leibniz in the 1650s.

[6] He actually made a small mistake, but he **could** have evaluated it to four decimal places.

[7] Remember that $E[X] = np$ and $\mathrm{Var}(X) = np(1-p)$ for a binomial.

Step 2. *When n is large*, de Moivre could also show that

$$P(X = n) = \binom{2n}{n} / 2^{2n} \approx \frac{1}{\sqrt{cn}},$$

where $c \in \mathbb{R}$ is some specific constant. At first he used an approximation of this constant but then his friend James Stirling stepped in to show that the constant is exactly equal to π![8] In other words, Stirling showed that

$$P(X = n) = \binom{2n}{n} / 2^{2n} \approx \frac{1}{\sqrt{\pi n}}.$$

///

By putting these two steps together we obtain the approximation

$$P(X = n + \ell) \approx \frac{1}{\sqrt{\pi n}}\, e^{-\ell^2/n},$$

which is valid when n is large and ℓ/n is small. The great advantage of this expression is that it is defined not just for whole numbers but *for all real numbers ℓ and n.* Going back to the case $2n = 3600$, we obtain the approximation

$$P(X = 1800 + \ell) \approx \frac{1}{\sqrt{1800\pi}}\, e^{-\ell^2/1800},$$

which is valid when the ratio $\ell/1800$ is small.[9] In particular, since $30/1800$ is rather small, de Moivre obtained a rather good estimate for $P(1700 \leqslant X \leqslant 1830)$ by integrating the function between -30 and 30. We can obtain an even better estimate by using the "continuity correction," i.e., by integrating from -30.5 to 30.5:

$$P(1770 \leqslant X \leqslant 1830) = \sum_{\ell=-30}^{30} \binom{3600}{1800+\ell} / 2^{3600} \approx \int_{-30.5}^{30.5} \frac{1}{\sqrt{1800\pi}}\, e^{-x^2/1800}\, dx \approx 69.07\%.$$

The integral on the right may not look any easier than the sum on the left. However, de Moivre knew how to expand the function $e^{-x^2/1800}$ as a power series in x and then he integrated this series term by term to get a convergent sum. It turns out that this sum converges so rapidly that only a few terms are needed to get a good approximation.

Many years later (around 1810) Pierre-Simon Laplace brough de Moivre's work to maturity by extending to the case of a general binomial.

The de Moivre-Laplace Theorem. Let X be a binomial random variable with parameters n and p. If the ratio k/np is *close to* 1, and if the numbers np and $n(1-p)$ are *both large*, then we have the following approximation:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-1} \approx \frac{1}{\sqrt{2\pi np(1-p)}}\, e^{-(k-np)^2/2np(1-p)}.$$

[8] No one was expecting that.

[9] How small? Never mind. We won't be concerned with fine details like that.

To simplify this expression a bit we should observe that $\mu = np$ and $\sigma^2 = np(1-p)$ are just the mean and variance of the binomial X. Thus we can also write

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-1} \approx \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(k-\mu)^2/2\sigma^2}.$$

///

I'm sorry that the difficulty of the math went up significantly in this lecture. The details of de Moivre's and Stirling's approximations are not so imporant for us, but the strange formula

$$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(k-\mu)^2/2\sigma^2}$$

that appears in the de Moivre-Laplace theorem will be **very** important.[10] This is the most important formula in statistics and it will be our main concern for the rest of the course. It is a bit messy but we're willing to put up with the mess because it is so useful.

Nov 16 and Thanksgiving Break

Last time I tried to motivate the following definition.

Definition of Normal Distribution. Let X be a continuous random variable. We say that X has a *normal distribution with parameters μ and σ^2* if its pdf is given by the formula

$$n(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}.$$

You might also see this written in the equivalent form

$$n(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

We will use the shorthand

$$X \sim N(\mu, \sigma^2)$$

to indicate that the random variable X has pdf given by $f_X(x) = n(x; \mu, \sigma^2)$. ///

The expression $n(x; \mu, \sigma^2)$ is meant to indicate that x is a variable and μ, σ^2 are constants. By treating this as a function of x we can see that

$$n(x; \mu, \sigma^2) \to 0 \qquad \text{as} \qquad x \to \pm\infty.$$

By computing the first two derivatives[11] one can also show that the graph of $n(x; \mu, \sigma^2)$

[10] You should memorize it.

[11] I'll let you do this on HW5.

- has a global maximum at $x = \mu$, and
- has inflection points at $x = \mu \pm \sigma$.

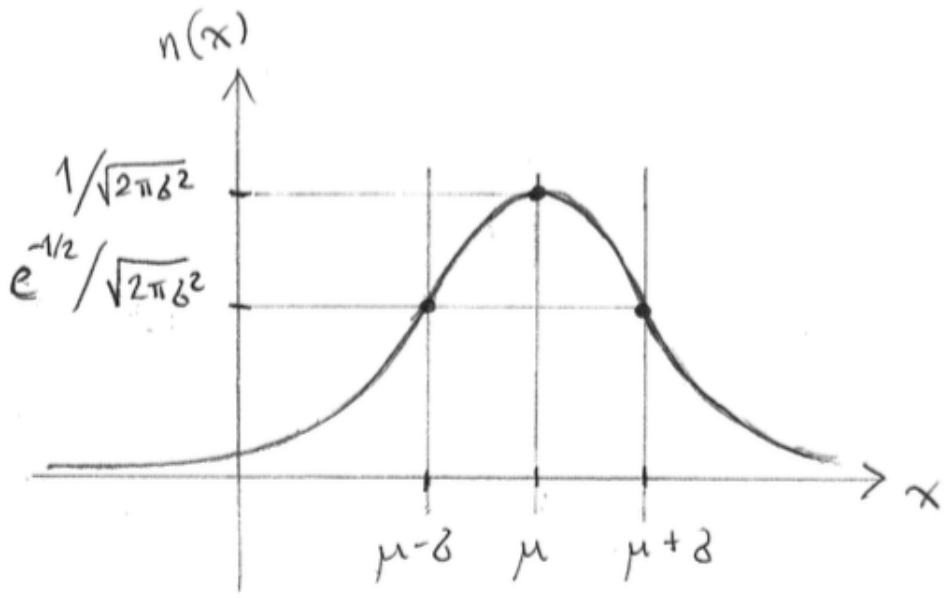
The heights of the maximum and the inflection points are

$$n(\mu; \mu, \sigma^2) = 1/\sqrt{2\pi\sigma^2} \qquad \text{and} \qquad n(\mu \pm \sigma; \mu, \sigma^2) = e^{-1/2}/\sqrt{2\pi\sigma^2},$$

which implies that the inflection points are always about 60% as high at the maximum:

$$\frac{n(\mu \pm \sigma; \mu, \sigma^2)}{n(\mu; \mu, \sigma^2)} = e^{-1/2} \approx 60.65\%.$$

In summary, the graph of $n(x; \mu, \sigma^2)$ looks like a "bell curve."



By calling this the "normal distribution" we have implicitly assumed that the total area under the curve is 1, i.e., that for any parameters μ and σ^2 we have

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}\, dx = \int_{-\infty}^{\infty} n(x; \mu, \sigma^2)\, dx = 1.$$

But why is this true? If it's true for any values of μ and σ^2 then it must be true for the special values $\mu = 0$ and $\sigma^2 = 1$. In this case the desired formula is

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

$$\int_{-\infty}^{\infty} e^{-x^2/2}\,dx = \sqrt{2\pi},$$

which we can clean up a bit by substituting $u^2 = x^2/2$ (and hence $dx = \sqrt{2}\,du$) to get

$$\boxed{\int_{-\infty}^{\infty} e^{-u^2}\,du = \sqrt{\pi}.}$$

Proving this boxed formula was the hardest piece of the de Moivre-Laplace theorem, the piece that was contributed by de Moivre's friend James Stirling. For this reason I'll call it *Stirling's formula*. It cannot be proved[12] with methods from Calc I and II, so we'll just take it as a basic fact of nature.

And what is the meaning of the parameters μ and σ^2? By using these particular symbols we are strongly implying that these parameters are the **mean** and **variance** of the distribution. Indeed, by using Stirling's formula and some integration by parts, one could show that

$$E[X] = \int_{-\infty}^{\infty} x \cdot n(x;\mu,\sigma^2)\,dx = \mu$$

and

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot n(x;\mu,\sigma^2)\,dx = \sigma^2.$$

But the computation is not very fun, and this isn't a Calculus class, so we won't bother.

In summary, we have an infinite family of normal distributions, one for each choice of mean μ and variance σ^2. The definition and the basic properties of this distributions are a bit tricky, so why do we bother with them? There are two reasons.

Reason 1. Normal distributions are everywhere.[13]

The de Moivre-Laplace theorem tells us that flipping many coins (or flipping one coin many times) gives an approximately normal distribution for the number of heads. To be specific, suppose that $X_1, X_2, \ldots, X_n$ is a sequence of **independent** Bernoulli random variables, each with expected value $E[X_i] = p$. Then the sum $X = X_1 + X_2 + \cdots + X_n$ is a binomial random variable with pmf

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

and we have seen many times[14] that $E[X] = np$ and $\mathrm{Var}(X) = np(1-p)$.

[12] The easiest proof uses a trick from multivariable Calculus. You can find it on Google if you're curious.

[13] That's why we call them normal.

[14] For example on Problem 1 of Exam 2.

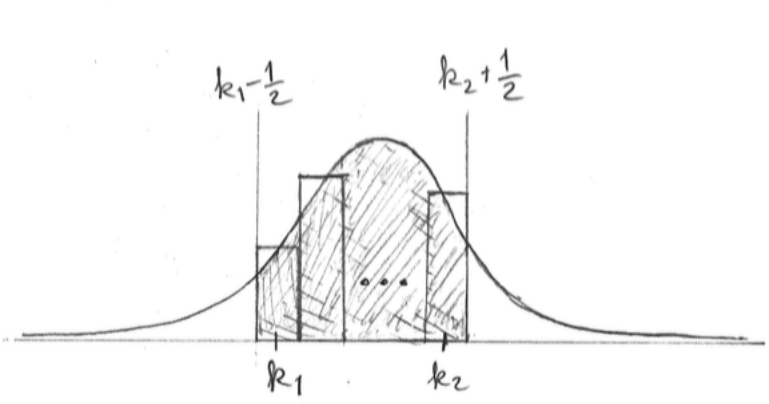
For small values of n we can compute binomial probabilities by hand. However, for large values of n we need some method of approximation. If np and $n(1-p)$ are both large enough (the rule of thumb is $np \geqslant 10$ and $n(1-p) \geqslant 10$) then the de Moivre-Laplace theorem tells us that

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \approx n(k; np, np(1-p)),$$

and this approximation is particularly good for values of k near the mean np. In other words, the binomial random variable X is approximately normal with parameters $\mu = np$ and $\sigma^2 = np(1-p)$. It follows that for any whole numbers k_1 and k_2 the probability of getting between k_1 and k_2 heads (inclusive) is approximated by an area under a normal curve:

$$P(k_1 \leqslant X \leqslant k_2) = \sum_{k=k_1}^{k_2} P(X = k) \approx \int_{k_1 - \frac{1}{2}}^{k_2 + \frac{1}{2}} n(x; np, np(1-p))\, dx.$$

We will compute some examples below. For now, here is a picture:



But that's not all. Laplace took this analysis even farther and showed that **any** sum of identical random variables is approximately normal. His "central limit theorem" is often referred to as the fundamental theorem of statistics.

The Central Limit Theorem. Let X be *any random variable whatsoever* with mean $E[X] = \mu$ and variance $\mathrm{Var}(X) = \sigma^2$. We can think of X as the result of some scientific measurement. If we perform the same measurement many times then we obtain a sequence of random variables, which we call a *sample*:

$$\begin{aligned} X_1 &= \text{result of the 1st measurement,} \\ X_2 &= \text{result of the 2nd measurement,} \\ &\vdots \\ X_n &= \text{result of the } n\text{th measurement.} \end{aligned}$$

We will assume that the random variables X_i are *identically distributed*, i.e., the underlying distribution X does not change between measurements, and we will assume that the measurements are *independent*, i.e., the thing being measured is not altered by our measuring procedure. We define the *sample mean* as the average of the n measurements:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

By linearity, the expected value of the sample mean is the same as the mean of the underlying distribution:

$$E[\overline{X}] = \frac{E[X_1] + E[X_2] + \cdots + E[X_n]}{n} = \frac{\mu + \mu + \cdots + \mu}{n} = \frac{n\mu}{n} = \mu.$$

This suggests that we can use the *sample mean* $\overline{X}$ (which we know) as an estimate for the underlying *population mean* μ (which we don't know). But how accurate is this estimate? Our intuition suggests that taking more observations will lead to more accuracy. Quantitatively, we can use the **independence** of the observations to show that

$$\begin{aligned}
\mathrm{Var}(\overline{X}) &= \mathrm{Var}(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n) \\
&= \frac{1}{n^2}\mathrm{Var}(X_1) + \frac{1}{n^2}\mathrm{Var}(X_2) + \cdots + \frac{1}{n^2}\mathrm{Var}(X_n) \\
&= \frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 \\
&= \frac{n\sigma^2}{n^2} \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

As the number n of observations grows, the variance of the average $\mathrm{Var}(\overline{X}) = \sigma^2/n$ goes to zero, which means that $\overline{X}$ is very likely to be close to μ. This phenomenon was first observed by Jacob Bernoulli, who called it the *law of large numbers*.[15]

In summary: If X is the result of some fixed experiment, then the average of n independent observations of this experiment has mean and variance given by

$$E[\overline{X}] = E[X] \qquad \text{and} \qquad \mathrm{Var}(\overline{X}) = \mathrm{Var}(X)/n.$$

This tells us that the known $\overline{X}$ is a good estimate for the unknown μ. But suppose we want to go further and estimate the probability that the unknown number μ lies within a certain fixed range. In that case we need to know the actual distribution of $\overline{X}$, and this is what Laplace's Central Limit Theorem tells us.

[15]The law of large numbers is the philosophical justification for the use of the expected value in statistics. That is, if you observe the random variable X many times, then on average you expect to get the expected value $E[X]$. That's reassuring.

Let X be any random variable and let $\overline{X}$ be the **average** *of n independent observations of X. If n is large, then the distribution of the sample mean is approximately normal with mean $E[X]$ and variance $\mathrm{Var}(X)/n$:*

$$\overline{X} \approx N\left(E[X], \mathrm{Var}(X)/n\right).$$

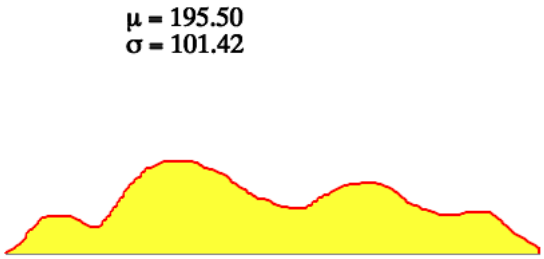
Equivalently, the **sum** *of the n observations, i.e., $n \cdot \overline{X}$, is approximately normal with mean $n \cdot E[X]$ and variance $n \cdot \mathrm{Var}(X)$:*

$$n \cdot \overline{X} \approx N\left(nE[X], n\mathrm{Var}(X)\right).$$

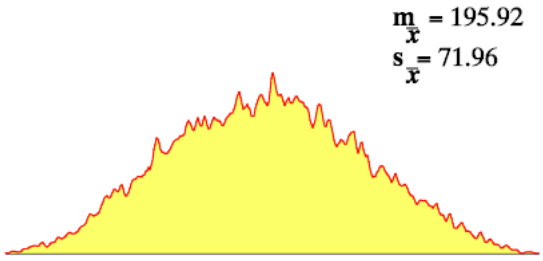
In the special case that X is a Bernoulli random variable with $E[X] = p$ and $\mathrm{Var}(X) = p(1-p)$ then the **sum** of n observations is binomial with parameters n and p, so we recover the de Moivre-Laplace theorem:

$$\text{binomial } n \cdot \overline{X} \approx N\left(nE[X], n\mathrm{Var}(X)\right) = N(np, np(1-p)).$$

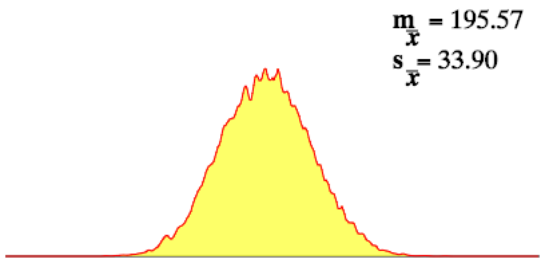
To illustrate the general idea, here is the pdf for some random variable X that I drew by hand. The underlying mean and standard deviation are $\mu = 195.5$ and $\sigma = 101.42$:



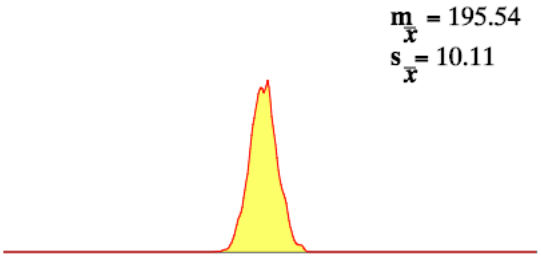
Here is the pdf for the average of 2 random samples from X:



Here is the pdf for the average of 9 random samples from X:



And here is the pdf for the average of 100 random samples from X:



We observe that the average of 100 random samples has a distribution that approximately normal with mean $\mu = 195.5$ and variance $\sigma^2/100$, i.e., standard deviation $\sigma/10 = 10.14$.[16]

Reason 2. Normal distributions have good "stability" properties.

We have seen that normal distributions occur whenever we take the average or the sum of many independent observations. That is the main reason why we care about normal random variables.

The other reason that normal distributions are so popular is because they have nice "stability" properties that allow us to work with them. This was especially important in the days before electronic computers.

Stability Properties of Normal Distributions. Let X, Y be normal random variables

$$X \sim N(\mu_X, \sigma_X^2) \qquad \text{and} \qquad Y \sim N(\mu_Y, \sigma_Y^2)$$

and let $\alpha, \beta \in \mathbb{R}$ be any constants. In this case the random variables $\alpha X + \beta$ and $\alpha X + \beta Y$ are also normal. That is, we have

$$\alpha X + \beta \sim N(\alpha\mu_X + \beta, \alpha^2\sigma_X^2),$$
$$\alpha X + \beta Y \sim N(\alpha\mu_X + \beta\mu_Y, \alpha^2\sigma_X^2 + \beta^2\sigma_Y^2).$$

You will prove the first of these properties on HW5. ///

The main application of these properties is that we can reduce any computation with normal distributions into a certain "standard" form.

Standardization of a Normal Distribution. Let $X \sim N(\mu, \sigma^2)$ be any normal random variable. Then we define its *standardization* by

$$Z = \frac{X - \mu}{\sigma}.$$

[16] The numbers $m_{\bar{x}}$ and $s_{\bar{x}}$ in the pictures are not the theoretical mean and standard deviation of $\overline{X}$. Instead, they are approximations that are generated by some random process. That's why they don't exactly match our predictions. I generated the pictures at this webpage: `http://www.ltcconline.net/greenl/java/Statistics/clt/cltsimulation.html`

By the first stability property we know that $Z = \frac{1}{\sigma}X - \frac{\mu}{\sigma}$ is normal. Furthermore, we can easily compute that

$$E[Z] = E\left[\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right] = \frac{1}{\sigma}E[X] - \frac{\mu}{\sigma} = \frac{1}{\sigma}\mu - \frac{\mu}{\sigma} = 0$$

and

$$E[Z^2] = E\left[\frac{(X-\mu)^2}{\sigma^2}\right] = \frac{1}{\sigma^2}E\left[(X-\mu)^2\right] = \frac{1}{\sigma^2}\sigma^2 = 1.$$

Hence we have $E[Z] = 0$ and $\mathrm{Var}(Z) = E[Z^2] - E[Z]^2 = 1 - 0^2 = 1$. In this case we say that Z has a *standard normal distribution*: $Z \sim N(0,1)$.[17]

In summary, we see that

$$\boxed{X \sim N(\mu, \sigma^2) \quad \Longleftrightarrow \quad Z = \frac{X-\mu}{\sigma} \sim N(0,1).}$$

///

Here's how we will apply this idea. Suppose that $X \sim N(\mu, \sigma^2)$ is any normally distributed random variable with mean μ and variance σ^2, so that $Z = (X-\mu)/\sigma$ has a standard normal distribution. Then for any real numbers $a \leqslant b$ we have

$$\begin{aligned}
\int_a^b n(x;\mu,\sigma^2)\,dx &= P(a \leqslant X \leqslant b) \\
&= P\left(a - \mu \leqslant X - \mu \leqslant b - \mu\right) \\
&= P\left(\frac{a-\mu}{\sigma} \leqslant \frac{X-\mu}{\sigma} \leqslant \frac{b-\mu}{\sigma}\right) \\
&= P\left(\frac{a-\mu}{\sigma} \leqslant Z \leqslant \frac{b-\mu}{\sigma}\right) = \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} n(x;0,1)\,dx.
\end{aligned}$$

We have reduced the problem of computing areas under a normal curve to the problem of computing areas under a **standard** normal curve. This is helpful because now instead of infinitely many different normal curves we only have to understand one of them. Unfortunately, this is still a hard problem.

The Bad News. To compute the area under a standard normal curve, we need to find an antiderivative for the density function

$$n(x;0,1) = \frac{1}{\sqrt{2\pi}}\,e^{-x^2/2}.$$

This antiderivative certainly *exists*. However, the bad news is that it does not have a formula that can be expressed in terms of any functions that we know (trigonometric, exponential,

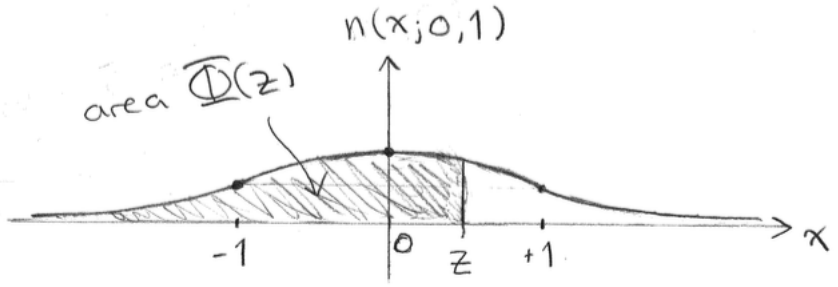
[17] The letter "z" is always used for the standard normal distribution. I have no idea why.

logarithmic, etc.). Thus our only choice is to give the antiderivative a new name. We will follow the textbook and call it $\Phi(z)$. By the Fundamental Theorem of Calculus we can define the antiderivative as the area under the curve from $-\infty$ to any real number $z \in \mathbb{R}$:

$$\Phi(z) = \int_{-\infty}^{z} n(x;0,1) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx.$$

If $Z \sim N(0,1)$ is any standard normal random variable then we also observe that $\Phi(z) = P(Z \leqslant z)$. This tells us that $\Phi(z) \to 0$ as $z \to -\infty$ and $\Phi(z) \to 1$ as $z \to \infty$. In probability and statistics books you will often see $\Phi(z)$ referred to as the *cumulative density function (cdf)* of the variable Z.

Here is a picture of the pdf $n(x;0,1)$ and the cdf $\Phi(z)$ of a standard normal, drawn to scale:



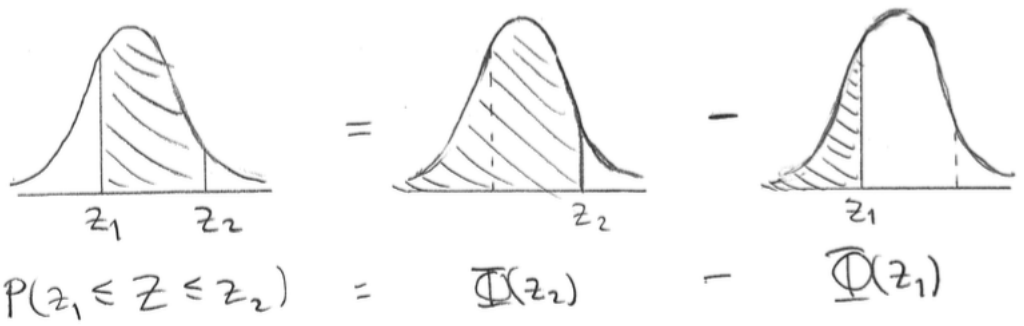
As you see, the pdf of a standard normal distribution is rather flat. For this reason we usually won't draw it to scale. ///

Fortunately, the difficult problem of computing $\Phi(z)$ has been thoroughly studied and we can look up the answers in the back of any statistics textbook.

The Good News. By using the Fundamental Theorem of Calculus we can express any area under the standard normal curve $n(x;0,1)$ in terms of its antiderivative $\Phi(z)$. Let Z be a standard normal random variable. Then for any real numbers $z_1 \leqslant z_2$ we have

$$P(z_1 \leqslant Z \leqslant z_2) = \int_{z_1}^{z_2} n(x;0,1)\, dx = \Phi(z_2) - \Phi(z_1).$$

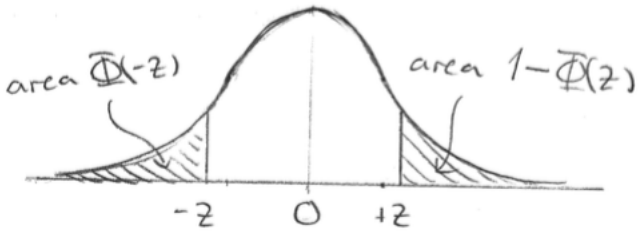
Here is a picture (not drawn to scale):



Furthermore, since the standard normal distribution is symmetric about zero we only need to know the values of $\Phi(z)$ when z is **positive**. Indeed, for any positive $z \in \mathbb{R}$ we observe that

$$\Phi(-z) = P(Z \leqslant -z) = P(Z \geqslant z) = 1 - \Phi(z).$$

Here is a picture (again, not to scale):



The good news is that a table of values for $\Phi(z)$ with $z \geqslant 0$ can be found in the back of any statistics textbook. In our textbook it's on page 494. ///

That was a lot of theory for one lecture so let me end with a couple of examples. After Thanksgiving we'll spend more time on applications.

Continuous Example. Suppose that X is normally distributed with mean $\mu = 6$ and variance $\sigma^2 = 25$, hence standard deviation $\sigma = 5$. Compute the probability that X is within one standard deviation of its mean:

$$P\left(|X - 6| < 5\right) = ?$$

Solution: Let $Z = (X - \mu)/\sigma = (X - 6)/5$ be the standardization, which has a standard normal distribution. Then we have

$$P\left(|X - 6| < 5\right) = P(-5 < X - 6 < 5)$$

$$
\begin{aligned}
&= P\left(\frac{-5}{5} < \frac{X-6}{5} < \frac{5}{5}\right) \\
&= P\left(-1 < Z < 1\right) \\
&= \Phi(1) - \Phi(-1) \\
&= \Phi(1) - [1 - \Phi(1)] \\
&= 2 \cdot \Phi(1) - 1 \\
&\approx 2(0.8413) - 1 \\
&= 0.6826 = 68.26\%.
\end{aligned}
$$

More generally, for any normal random variable $X \sim N(\mu, \sigma^2)$ we find that

$$
P\left(|X - \mu| < \sigma\right) = P\left(-1 < \frac{X-\mu}{\sigma} < 1\right) = \Phi(1) - \Phi(-1) = 2 \cdot \Phi(1) - 1 \approx 68.26\%.
$$

and also that

$$
\begin{aligned}
P\left(|X - \mu| < 2\sigma\right) &= 2 \cdot \Phi(2) - 1 \approx 95.44\%, \\
P\left(|X - \mu| < 3\sigma\right) &= 2 \cdot \Phi(3) - 1 \approx 99.74\%.
\end{aligned}
$$

It is useful to remember this *68–95–99.7 rule* because normal distributions are so common.

Binomial Approximation Example. Suppose we flip a fair coin 3600 times and let X be the number of heads that we get. Compute the probability that we get between 1790 and 1815 heads (inclusive):

$$
P(1790 \leqslant X \leqslant 1815) = ?
$$

Solution: Since X is a binomial random variable with $n = 3600$ and $p = 1/2$ we know that

$$
\mu = np = 1800 \qquad \text{and} \qquad \sigma = \sqrt{np(1-p)} = 30.
$$

By the de Moivre-Laplace Theorem (which is a special case of the Central Limit Theorem) we know that X is approximately normal with the same mean and standard deviation, and hence that $(X - 1800)/30$ is approximately a standard normal random variable. Thus we have

$$
\begin{aligned}
P(1790 \leqslant X \leqslant 1815) &= P(-10 \leqslant X - 1800 \leqslant 15) \\
&= P\left(\frac{-10}{30} \leqslant \frac{X-1800}{30} \leqslant \frac{15}{30}\right) \\
&= P\left(\frac{-1}{3} \leqslant \frac{X-1800}{30} \leqslant \frac{1}{2}\right) \\
&\approx \Phi(1/2) - \Phi(-1/3) \\
&= \Phi(1/2) - [1 - \Phi(1/3)] \\
&\approx (0.6915) - [1 - (0.6293)] = 0.3208 = 32.08\%.
\end{aligned}
$$

We get an even better approximation if we remember to apply the "continuity correction." Let $X' \sim N(1800, 30^2)$ be a (continuous) normal variable with the same mean and standard deviation as the discrete variable X. Then we have

$$\begin{aligned} P(1790 \leqslant X \leqslant 1815) &\approx P(1789.5 \leqslant X' \leqslant 1815.5) \\ &= P(-10.5 \leqslant X' - 1800 \leqslant 15.5) \\ &= P\left(\frac{-10.5}{30} \leqslant \frac{X' - 1800}{30} \leqslant \frac{15.5}{30}\right) \\ &\approx P\left(-0.35 \leqslant \frac{X' - 1800}{30} \leqslant 0.52\right) \\ &= \Phi(0.52) - \Phi(-0.35) \\ &= \Phi(0.52) - [1 - \Phi(0.35)] \\ &\approx (0.6985) - [1 - (0.6368)] = 0.3353 = 33.53\%. \end{aligned}$$

For comparison, my laptop is powerful enough to compute the **exact** answer:

$$P(1790 \leqslant X \leqslant 1815) \approx 33.41\%.$$

This shows that the "continuity correction" gives a better result.

Nov 28

We are ready to consider our first official statistics problem.

Statistics Problem Part I. We have a coin[18] with $P(H) = p$ where the value of p is unknown to us. We want to perform an experiment to estimate the value of p. ///

It is pretty clear what our experiment should be. We will flip the coin n times and let X be the number of heads that we get. Then we expect that the *sample mean* $\overline{X} = X/n$ will be close to the true value of p. In this case we will write

$$\hat{p} = \overline{X}$$

and say that we are using the random variable $\overline{X}$ as an *estimator* for the unknown *parameter* p. Since X is a binomial random variable with parameters n, p we recall that $E[X] = np$. Hence

$$E[\hat{p}] = E[\overline{X}] = E\left[\frac{1}{n} \cdot X\right] = \frac{1}{n} \cdot E[X] = \frac{1}{n} \cdot np = p.$$

In this case we say that $\overline{X}$ is an *unbiased estimator* for p. That is, if we perform this experiment many times then **on average** we will obtain the correct answer.

[18] Not necessarily a literal coin.

Real World Example. Consider a population of voters and suppose that p is the proportion of these voters who plan to vote for a certain "thing."[19] For simplicity we will model each voter as an identical coin flip where "heads" means "yes," so that $P(H) = p$. In order to estimate the true value of p we polled $n = 400$ voters and found that $X = 136$ plan to vote yes. Therefore our estimate for p is

$$\hat{p} = \overline{X} = \frac{X}{n} = \frac{136}{400} = 34\%.$$

In other words, we estimate that 34% of the total population plans to vote yes. ///

Well, that's great but we can't just report this "statistic" without giving some measure of how accurate we think it is.

Statistics Problem Part II. Consider a coin with an unknown parameter of $P(H) = p$. In order to estimate this parameter we flip the coin n times and let X be the number of heads that we get. Then we report the number

$$\hat{p} = \overline{X} = X/n$$

as our guess for the true value of p. But how **confident** are we in this guess? ///

This is a fairly ambiguous question and there are many different points of view. Today we'll discuss the notion of a "confidence interval," which is the most common way to approach the problem.

The Idea of a Confidence Interval. We have a constant unknown parameter p and a random variable $\hat{p}$ which is an unbiased estimator for p, i.e., $E[\hat{p}] = p$. We know that our guess might be wrong so we want to find some estimate of the error. In the best case scenario we are looking for a number $e \in \mathbb{R}$ so that the true value of p is **guaranteed** to lie between the numbers $\hat{p} - e$ and $\hat{p} + e$. But, sadly, this "absolute certainty" is impossible.

Therefore we are willing to settle for "moral certainty." For example, if 95% certainty is good enough then we will search for a number e with the property that

$$P(\hat{p} - e < p < \hat{p} + e) = 95\%.$$

This means that the true value of p will lie between $\hat{p} - e$ and $\hat{p} + e$ in 95 out of every 100 runs of the experiment. ///

Let's try to compute such a "95% confidence interval" for our polling example. Recall that X is a binomial random variable with known parameter $n = 400$ and unknown parameter p. We will use the sample mean $\hat{p} = X/400$ as an estimator for p.

[19] I'm not going to get all political here.

Since the mean and variance of the binomial random variable X are given by $E[X] = 400p$ and $\mathrm{Var}(X) = 400p(1-p)$, and since the number 400 is rather large, the Central Limit Theorem tells us that $(X - 400p)/\sqrt{400p(1-p)}$ is approximately standard normal. In order to express this in term of $\hat{p}$ we write

$$N(0,1) \approx \frac{X - 400p}{\sqrt{400p(1-p)}} = \frac{X/400 - p}{\sqrt{p(1-p)/400}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/400}} = \frac{20(\hat{p} - p)}{\sqrt{p(1-p)}}.$$

If Z is a true standard normal variable then we can look up in our table that

$$P(-1.96 < Z < 1.96) = 95\%.$$

Hence in our case we obtain an approximation[20]

$$\begin{aligned}
95\% &\approx P\left(-1.96 < \frac{20(\hat{p} - p)}{\sqrt{p(1-p)}} < 1.96\right) \\
&= P\left(-1.96 \cdot \frac{\sqrt{p(1-p)}}{20} < \hat{p} - p < 1.96 \cdot \frac{\sqrt{p(1-p)}}{20}\right) \\
&= P\left(-1.96 \cdot \frac{\sqrt{p(1-p)}}{20} < p - \hat{p} < 1.96 \cdot \frac{\sqrt{p(1-p)}}{20}\right) \\
&= P\left(\hat{p} - 1.96 \cdot \frac{\sqrt{p(1-p)}}{20} < p < \hat{p} + 1.96 \cdot \frac{\sqrt{p(1-p)}}{20}\right) \\
&= P\left(\hat{p} - e < p < \hat{p} + e\right),
\end{aligned}$$

where the error is given by

$$e = 1.96 \cdot \frac{\sqrt{p(1-p)}}{20} = 0.098 \cdot \sqrt{p(1-p)}.$$

Here's the good news and the bad news:

- The **good news** is that we have found an approximate 95% confidence interval for p. That is, in approximately 95% of the runs of this experiment, the true value of p will lie between $\hat{p} - e$ and $\hat{p} + e$.
- The **bad news** is that our formula for e depends on the **unknown** parameter p which we are trying to estimate. Therefore our result is completely useless.

Or is it? We can't just give up, so we need to find an approximate value for the error e. The idea here is obvious, even if it is mathematically dubious:

[20] You might wonder we don't use a continuity correction here. Maybe we should, but then our answer won't match the textbook answer.

In the formula for e we will replace the unknown p by the known $\hat{p}$. Since 400 is rather large the variance $\mathrm{Var}(\hat{p}) = p(1-p)/400$ is rather small, so it is probably okay to assume that $\hat{p} \approx p$. Isn't this reasoning a bit circular? Yes it is, but hopefully it works anyway. We can always try to improve it later.

In other words, we assume that the following approximation holds:

$$P\left(\hat{p} - 1.96 \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{20} < p < \hat{p} + 1.96 \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{20}\right) \approx 95\%.$$

Finally, by plugging in the experimental value $\hat{p} = X/400 = 136/400 = 0.34$ we obtain

$$P\left(0.34 - 1.96 \cdot \frac{\sqrt{(0.34)(0.66)}}{20} < p < 0.34 + 1.96 \cdot \frac{\sqrt{(0.34)(0.66)}}{20}\right) \approx 95\%$$

$$P\left(0.34 - 0.046 < p < 0.34 + 0.046\right) \approx 95\%$$

In summary: A poll of 400 people finds that 136 plan to vote "yes." If p is the true proportion of "yes" voters in the whole population then we can report to our boss that

$$\begin{aligned} 34\% - 4.6\% &< p < 34\% + 4.6\%, \\ 29.4\% &< p < 38.6\%, \end{aligned}$$

with a confidence of approximately 95%.

Here's the general story.

Confidence Intervals for Proportions. Suppose we have a coin where the probability of heads $P(H) = p$ is unknown. To estimate the parameter p we flip the coin n times and let X be the number of heads we get. Then $\hat{p} = X/n$ is an unbiased estimator for p.

Suppose that $Z \sim N(0,1)$ and let $z_{\alpha/2}$ be the unique number satisfying

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha.$$

For example, we have seen that $z_{0.05/2} = z_{0.025} = 1.96$. Then we have the following formulas for $(1-\alpha)100\%$ confidence intervals:

(1) If n is **large** then we assume that $p \approx \hat{p}$ and hence

$$P\left(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 1 - \alpha.$$

(2) If n is **small** then instead of $\hat{p} \pm z_{\alpha/2} \cdot \sqrt{p(1-p)/n}$ we use the more accurate[21] interval

$$\frac{\hat{p} + z_{\alpha/2}^2/(2n) \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n + z_{\alpha/2}^2/(4n^2)}}{1 + z_{\alpha/2}^2/n}.$$

These intervals are applicable as long as the true proportion p is not too close to 0 or 1. When this is not the case our textbook suggests the following rule:

(3) If we have reason to believe that p is close to 0 or 1 then instead of the sample mean $\hat{p} = X/n$ we will use the strange estimator[22] $\tilde{p} = (X+2)/(n+4)$. If n is large then by reasoning similar to (1) we obtain

$$P\left(\tilde{p} - z_{\alpha/2} \cdot \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} < p < \tilde{p} + z_{\alpha/2} \cdot \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}\right) \approx 1 - \alpha.$$

If p is close to 0 or 1 and if n is small then I have no idea what to do.

And here's a final application.

Application (Hypothesis Testing[23]). Suppose we have a coin with an unknown parameter $p = P(H)$. We flip the coin 200 times and get heads 116 times. Is the coin fair?

Let H_0 be the event "the coin is fair," which we call our *null hypothesis*. In order to test this hypothesis we will compute an approximate 95% confidence interval for the unknown p and we will *reject* H_0 if the *null value* $p_0 = 50\%$ falls outside this interval. Otherwise we will *fail to reject* H_0.

Our estimator for p is $\hat{p} = 116/200 = 58\%$. Since $n = 200$ is relatively large we will use the most basic confidence interval, number (1) above. At the $(1-\alpha)100\% = 95\%$ of confidence we have

$$P\left(0.58 - 1.96 \cdot \sqrt{\frac{(0.58)(0.42)}{200}} < p < 0.58 + 1.96 \cdot \sqrt{\frac{(0.58)(0.42)}{200}}\right) \approx 95\%,$$

$$P\left(58\% - 6.84\% < p < 58\% + 6.84\%\right) \approx 95\%,$$

$$P\left(51.16\% < p < 64.84\%\right) \approx 95\%.$$

Since $p_0 = 50\%$ does not fall in the 95% confidence interval $[51.16\%, 64.84\%]$ for the unknown p, we *reject* the null hypothesis H_0. In other words, we conclude that the coin is **not fair**.

[21] Details omitted.

[22] Which is also biased.

[23] The hypothesis test in this example is a bit nonstandard. If we had more time I would use the jargon of Type I and Type II errors expressed with the letters α and β. However, there is no new mathematics involved in that stuff so you should be able to learn it quickly on your own.

Nov 30 and Dec 5

Last time we discussed confidence intervals for proportions. For example, these allow us to estimate the probability of "heads" for an unknown coin, or to estimate the proportion of "yes" voters in a certain population by performing a random poll. Mathematically we can model both of these experiments as taking independent samples of an underlying **Bernoulli distribution** with an unknown parameter p.

On ther other extreme, suppose we have a distribution that is approximately normal with unknown parameters μ and σ^2. In this case we would like to estimate the values of μ and σ^2 by taking independent samples from the distribution. That's what we'll do today. But first let me clarify some definitions from last time.

Definition of Confidence Intervals. Let X be a random variable depending on some fixed but unknown parameter θ. Let $X_1, X_2, \ldots, X_n$ be the results of n independent samples from X and let $\hat{\theta}$ be some random variable that is computed from the samples. We say that $\hat{\theta}$ is an *unbiased estimator* for θ if

$$E[\hat{\theta}] = \theta.$$

For this to be useful, we need to know how close (on average) this estimator is to the true value of θ. There are many ways to address this problem, the most popular of which is to compute a *confidence interval*, as follows. We say that $\hat{\theta} \pm e_{\alpha/2}$ is a **two-sided** $(1-\alpha)100\%$ **confidence interval for** θ if

$$P(\hat{\theta} - e_{\alpha/2} < \theta < \hat{\theta} + e_{\alpha/2}) = 1 - \alpha.$$

Please note here that the unknown parameter θ is **constant**, while the estimator $\hat{\theta}$ is a **random variable**. The equation above says that the true value θ will fall within the random interval $\hat{\theta} \pm e_{\alpha/2}$ for $(1-\alpha)100\%$ of the runs of this experiment.

If desired we can also consider **one-sided** $(1-\alpha)100\%$ **confidence intervals** given by

$$P(\hat{\theta} - e_\alpha < \theta) = 1 - \alpha,$$
$$P(\theta < \hat{\theta} + e_\alpha) = 1 - \alpha.$$

///

We assume that the estimator $\hat{\theta}$ is easy to compute. For example, it could be the sample mean $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$. Thus the real difficulty is to compute the error bounds e_α and $e_{\alpha/2}$ for various values of α. For this purpose it is convenient to make the following definition.

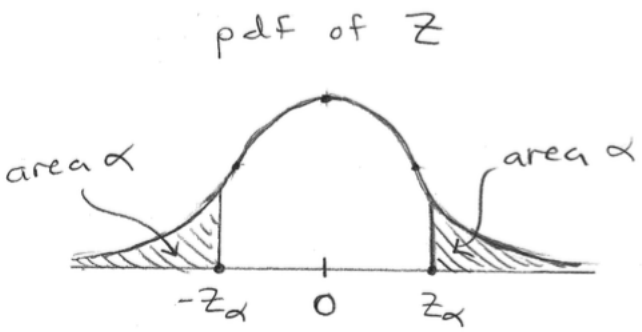
P-Values for the Standard Normal Distribution.

The "P" in the term "P-value" stands for "probability." However, since "P" is currently being used for other things (such as the probability of heads) we will use "α" instead.

Let $Z \sim N(0,1)$ be a standard normal random variable and let $0 < \alpha < 1$ be any possible probability value (i.e., P-value). Then there exists a unique number z_α with the property

$$\boxed{P(Z \geqslant z_\alpha) = \alpha.}$$

This is clearer if we look at a picture. Here, α is the area of the tail to the right of z_α:



By symmetry, the area of the left tail to the left of $-z_\alpha$ is also equal to α, so that

$$\alpha = P(Z \leqslant -z_\alpha) = P(Z < -z_\alpha) = 1 - P(Z \geqslant -z_\alpha)$$

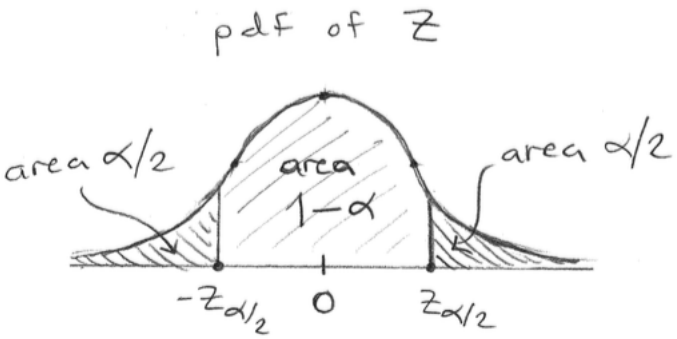
and hence $P(Z \geqslant -z_\alpha) = 1 - \alpha$. But $z_{1-\alpha}$ is by definition the unique number satisfying $P(Z \geqslant z_{1-\alpha}) = 1 - \alpha$, which implies that

$$\boxed{z_{1-\alpha} = -z_\alpha.}$$

Finally, let me note that

$$\boxed{P(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}) = 1 - \alpha,}$$

which can be seen in the following picture:



A table of the numbers z_α can be found on page 495 of our textbook, however this table is not really necessary. Indeed, because of the formula

$$\Phi(z_\alpha) = 1 - \alpha$$

we could find the same information by doing a reverse look-up in the table of $\Phi(z)$ values on page 494. I honestly have no idea why they include both tables. ///

We saw last time that the numbers z_α can be used to compute confidence intervals for the mean of an unknown Bernoulli distribution. Today we'll use them to compute confidence intervals for the mean of an unknown normal distribution.

Let $X_1, X_2, \ldots, X_n$ be independent samples from a normal distribution $N(\mu, \sigma^2)$. As usual, the sample mean $\overline{X} = (X_1 + \cdots + X_n)/n$ is an unbiased estimator[24] for the population mean μ. Therefore we will write

$$\hat{\mu} = \overline{X}.$$

When X was a Bernoulli, we knew from the Central Limit Theorem that $\overline{X}$ was approximately normal. Now the situation is even nicer.

The Sample Mean for Normal Distribution is Normal. Let $X \sim N(\mu, \sigma^2)$ and let $\overline{X} = (X_1 + \cdots + X_n)/n$ be the mean of n independent samples from X. Then $\overline{X}$ has an exactly normal distribution:

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Proof: The normality of $\overline{X}$ follows from the "stability" properties of normal distributions, of which you proved a special case on HW5. Thus we only need to check that $\overline{X}$ has the claimed mean and variance:[25]

$$E[\overline{X}] = \frac{1}{n}\left(E[X_1] + \cdots + E[X_n]\right) = \frac{1}{n}(\mu + \cdots + \mu) = \frac{1}{n} \cdot n\mu = \mu,$$
$$\mathrm{Var}(\overline{X}) = \frac{1}{n^2}\left(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)\right) = \frac{1}{n^2}(\sigma^2 + \cdots + \sigma^2) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

///

It follows from this (by the same "stability" properties) that the standardization of $\overline{X}$ has an exactly standard normal distribution:

$$\frac{\overline{X} - E[\overline{X}]}{\sqrt{\mathrm{Var}(\overline{X})}} = \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

[24] This is true for any distribution whatsoever, not only for the Bernoulli and normal distributions.

[25] We already did this when we discussed the Central Limit Theorem but sufficiently many of you probably don't remember.

Let's use this to find an exact $(1-\alpha)100\%$ confidence interval for the mean. By definition of the numbers z_α we have

$$\begin{aligned}
1-\alpha &= P\left(-z_{\alpha/2} < \frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) \\
&= P\left(-z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < \hat{\mu}-\mu < z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(-z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < \mu-\hat{\mu} < z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\hat{\mu}-z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < \mu < \hat{\mu}+z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right).
\end{aligned}$$

As with Bernoulli sampling, there is good news and bad news:

- The **good news** is that we have found an exact $(1-\alpha)100\%$ confidence interval for the mean μ of a normal distribution. That is, if we take n independent samples and compute the sample mean $\overline{X}$, then there is an exactly $(1-\alpha)100\%$ chance that the true mean μ lies in the interval $\overline{X} \pm z_{\alpha/2}\cdot\sigma/\sqrt{n}$.
- The **bad news** is that our formula for this confidence interval depends on the **unknown** parameter σ. (Indeed, if μ is unknown to us then why would we know σ?) Therefore our result is completely useless.

Or is it? Since it's not okay to give up, we must come up with some kind of estimator for the unknown variance σ^2. Our first guess might be to use the formula

$$V = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2.$$

This formula is not so bad. However, you will prove on HW6 that the expected value of the random variable V is

$$E[V] = \frac{n-1}{n}\cdot\sigma^2 \neq \sigma^2,$$

and hence V is a (slightly) **biased** estimator for the variance. In order to fix this situation we make the following definition.

Sample Variance and Sample Standard Deviation. Suppose that we take a random sample $X_1, X_2, \ldots, X_n$ from an underlying population with mean μ and variance σ^2. Then we define the *sample variance* by the formula

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

Please note that the strange random variable S^2 is related to the nice random variable V by the equation $S^2 = \frac{n}{n-1}\cdot V$ and your computation from HW6 gives

$$E[S^2] = E\left[\frac{n}{n-1}\cdot V\right] = \frac{n}{n-1}\cdot E[V] = \frac{\cancel{n}}{\cancel{n-1}}\cdot\frac{\cancel{n-1}}{\cancel{n}}\cdot\sigma^2 = \sigma^2.$$

We conclude that the sample variance S^2 is an **unbiased** estimator for the population variance σ^2. This is the reason that we use $n-1$ in the denominator instead of n.

As suggested by the notation, we define the *sample standard deviation* S as the positive square root of the sample variance:[26]

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

///

Now by using S as an estimator for the standard deviation σ we can finally compute confidence intervals for the mean of a normal distribution.

Confidence Intervals for the Mean of a Normal Distribution. Let $X_1, X_2, \ldots, X_n$ be independent samples from a normal population with mean μ and σ^2. Let $\overline{X}$ and S be the sample mean and sample standard deviation. Then we have the following $(1-\alpha)100\%$ confidence intervals for μ:

(1) If n is **large** then the sample standard deviation S is a good enough estimator for σ so that the random variable $(\overline{X} - \mu)/(S/\sqrt{n})$ has an approximately standard normal distribution:

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

Thus we obtain an approximate confidence interval for the unknown mean:

$$P\left(\overline{X} - z_{\alpha/2} \cdot \frac{S}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}\right) \approx 1 - \alpha.$$

(2) If n is **small** then the normal approximation from part (1) is not good enough. Instead we need to know the exact distribution of the standardized sample mean:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \text{ has a "}t\text{ distribution with } n-1 \text{ degrees of freedom."}$$

I don't want to get into the details of t distributions. I will simply note that for each whole number r there is a "t distribution with r degrees of freedom" which looks approximately like a normal curve. The corresponding critical values $t_\alpha(r)$ can be looked up in the table on page 465. Thus we obtain an exact confidence interval for the unknown mean:

$$P\left(\overline{X} - t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

[26] Recall that the expectation function does not preserve multiplication. That is, in general we have $E[XY] \neq E[X] \cdot E[Y]$. Therefore we have no reason to expect that $E[S]^2 = E[S^2]$, and in fact this equation is **false**. In other words, the sample standard deviation is a **biased** estimator for the population standard deviation. Oh well, we can't win every time.

These confidence intervals are still reasonable when the underlying population is only **approximately normal**. If the underlying population has a different shape then we will need different techniques.

I will illustrate these ideas with an example from the textbook.

Example 7.1-5. Let X be the amount of butterfat that a typical cow produces during a certain period of time.[27] We will assume that X has a normal distribution $N(\mu, \sigma^2)$. In order to estimate μ a farmer measured the butterfat production for $n = 20$ cows and obtained the the following independent observations $X_1, X_2, \ldots, X_{20}$:

481	537	513	583	453	510	570	500	457	555
618	327	350	643	499	421	505	637	599	392

From these 20 observations we compute the sample mean

$$\overline{X} = \frac{1}{20}\sum_{i=1}^{20} X_i = 507.5$$

and the sample standard deviation

$$S = \sqrt{\frac{1}{20-1}\sum_{i=1}^{20}(X_i - \overline{X})^2} = 89.75.$$

We will use $\overline{X} = 507.5$ as our point estimate for the population mean μ. But how accurate is this estimate?

If we knew σ exactly then we would know that $(\overline{X} - \mu)^2/(\sigma/\sqrt{20})$ is standard normal and we would proceed from there to compute an exact confidence interval:

$$P\left(507.5 - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{20}} < \mu < 507.5 + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{20}}\right) = 1 - \alpha.$$

However, since we don't know σ, we have to use the estimator $S = 89.75$. If the number of samples $n = 20$ were large we would still obtain an approximate confidence interval:

$$P\left(507.5 - z_{\alpha/2} \cdot \frac{89.75}{\sqrt{20}} < \mu < 507.5 + z_{\alpha/2} \cdot \frac{89.75}{\sqrt{20}}\right) \approx 1 - \alpha.$$

However, since $n = 20$ is relatively small we should be more precise and use the fact that $(\overline{X} - \mu)^2/(S/\sqrt{20})$ has a t distribution with $20 - 1 = 19$ degrees of freedom. Then we obtain an exact confidence interval:

$$P\left(507.5 - t_{\alpha/2}(19) \cdot \frac{89.75}{\sqrt{20}} < \mu < 507.5 + t_{\alpha/2}(19) \cdot \frac{89.75}{\sqrt{20}}\right) = 1 - \alpha.$$

[27] Irrelevant details omitted.

And it only remains to choose our desired level of confidence. For a $(1-\alpha)100\% = 90\%$ confidence interval we look in the table on page 496 to find $t_{\alpha/2}(19) = t_{0.05}(19) = 1.729$ and then we obtain

$$P\left(507.5 - 1.729 \cdot \frac{89.75}{\sqrt{20}} < \mu < 507.5 + 1.729 \cdot \frac{89.75}{\sqrt{20}}\right) = 90\%,$$
$$P\left(507.5 - 34.7 < \mu < 507.5 + 34.7\right) = 90\%,$$
$$P\left(472.8 < \mu < 542.2\right) = 90\%.$$

///

Closing Remarks: These results are only valid as long as the underlying distribution of butterfat X is (approximately) normal. There are methods that one could use to check this assumption (such as a *quantile-quantile plot*, see page 253 in the text). If it turns out that X is not normally distributed then the methods discussed here are not appropriate. Then we would have two options.

(1) Find a more accurate guess for the distribution shape of X and use exact information about such curves to compute confidence intervals for μ.

(2) Instead of looking for the mean μ we could investigate the *median*, or 50-*th percentile* $m = \pi_{0.5}$ of the population. There exist techniques to compute confidence intervals for the median (and other percentiles) that are independent of the shape of the underlying distribution of X. (See section 7.5 of the text.)

You will learn about these things if you go further into statistics.

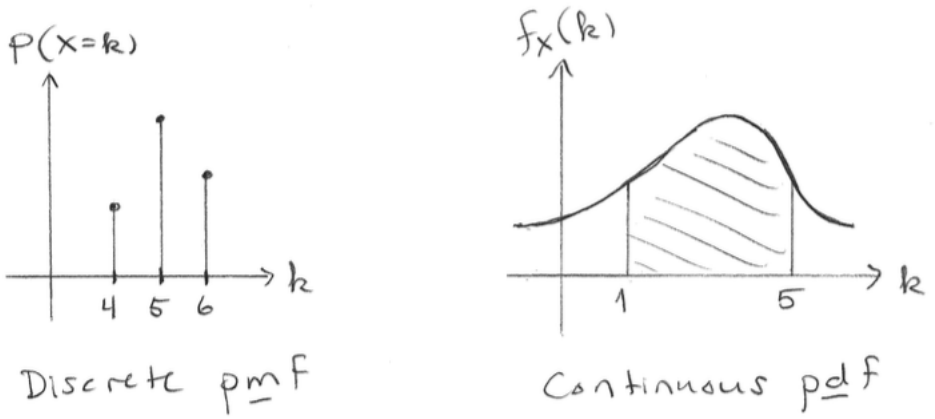
Dec 7

We discussed the solutions to HW6 and then I gave a review of topics for Exam 3. The only difference from Exams 1 and 2 is that you **will** be able to use scientific calculators (no wi-fi or cell signals). I will also provide tables of relevant statistical functions. Here's the review:

- **Discrete vs. Continuous Random Variables I.** A discrete random variable X has a probability *mass* function (pmf) defined by

$$f_X(k) = \begin{cases} P(X = k) & \text{if } k \text{ is an integer,} \\ 0 & \text{otherwise.} \end{cases}$$

  However, this won't work for continuous random variables. Indeed, if X is a continuous random variable then for any fixed number k we must have $P(X = k) = 0$. Instead, we define a continuous random variable in terms of a probability *density* function (pdf) f_X as in the following picture:



For the discrete random variable on the left we have

$$\begin{aligned}
P(4 \leqslant X \leqslant 6) &= P(X = 4) + P(X = 5) + P(X = 6) \\
P(4 < X \leqslant 6) &= P(X = 5) + P(X = 6) \\
P(4 \leqslant X < 6) &= P(X = 4) + P(X = 5) \\
P(4 < X < 6) &= P(X = 5).
\end{aligned}$$

For the continuous random variable on the right we have

$$P(1 \leqslant X \leqslant 5) = P(1 < X \leqslant 5) = P(1 \leqslant X < 5) = P(1 < X < 5) = \int_1^5 f_X(k)\, dk.$$

We say that a general function $f : \mathbb{R} \to \mathbb{R}$ is a pdf when it satisfies

$$f(x) \geqslant 0 \text{ for all } x \in \mathbb{R} \qquad \text{and} \qquad \int_{-\infty}^{\infty} f(x)\, dx = 1.$$

- **Expected Value and Variance.** Let $f_X : \mathbb{R} \to \mathbb{R}$ be the pdf of a continuous random variable X. Then we define the expected value by the formula

$$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx.$$

Just as in the discrete case, this integral represents the *center of mass* of the distribution. More generally, we define the rth moment of X by the formula

$$E[X^r] = \int_{-\infty}^{\infty} x^r \cdot f_X(x)\, dx.$$

As with the discrete case, the variance is defined as the expected distance between X and its mean $\mu = E[X]$. That is, we have

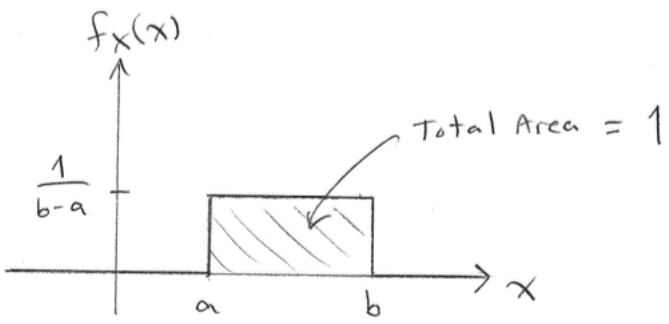
$$\text{Var}(X) = E[(X - \mu)^2]$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f_X(x)\,dx \\
&= \int_{-\infty}^{\infty} (x^2 - 2\mu x + x^2) \cdot f_X(x)\,dx \\
&= \left(\int_{-\infty}^{\infty} x^2 \cdot f_X(x)\,dx\right) - 2\mu\left(\int_{-\infty}^{\infty} x \cdot f_X(x)\,dx\right) + \mu^2\left(\int_{-\infty}^{\infty} f_X(x)\,dx\right) \\
&= E[X^2] - 2\mu \cdot E[X] + \mu^2 \cdot 1 \\
&= E[X^2] - 2\mu^2 + \mu^2 \\
&= E[X^2] - \mu^2 \\
&= E[X^2] - E[X]^2.
\end{aligned}
$$

- **Example: The Uniform Distribution.** The *uniform* distribution on a real interval $[a, b] \subseteq \mathbb{R}$ has pdf defined by

$$
f_X(x) = \begin{cases} 1/(b-a) & a \leqslant x \leqslant b, \\ 0 & \text{otherwise.} \end{cases}
$$

  This is a pdf because we have $f_X(x) \geqslant 0$ for all $x \in \mathbb{R}$ and the total area under the curve is 1. Here's a picture:



  You should practice the definitions by proving that

$$
E[X] = \frac{a+b}{2} \qquad \text{and} \qquad \mathrm{Var}(X) = \frac{(b-a)^2}{12}.
$$

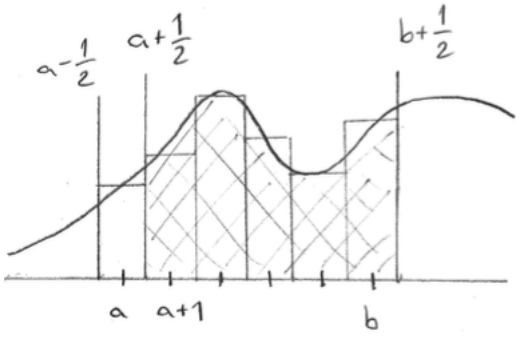
- **Discrete vs. Continuous Random Variables II.** Discrete and continuous random variables can be directly compared by looking at probability histograms. Let X be a discrete random variable with pmf $P(X = k)$ and let Y be a continuous random variable with pdf f_Y. Suppose that for all integers k we have

$$
P(X = k) \approx f_Y(k).
$$

Then for any integers $a \leqslant b$ we can approximate the probability $P(a \leqslant X \leqslant b)$ by the area under the graph of f_Y, as follows:

$$P(a \leqslant X \leqslant b) \approx \int_{a-1/2}^{b+1/2} f_Y(t)\,dt,$$
$$P(a < X \leqslant b) \approx \int_{a+1/2}^{b+1/2} f_Y(t)\,dt,$$
$$P(a \leqslant X < b) \approx \int_{a-1/2}^{b-1/2} f_Y(t)\,dt,$$
$$P(a < X < b) \approx \int_{a+1/2}^{b-1/2} f_Y(t)\,dt.$$

Here's a picture illustrating the second formula:



- **De Moivre-Laplace.** Let X be a (discrete) binomial random variable with parameters n and p. If np and $n(1-p)$ are both large, Abraham de Moivre (1730) and Pierre Laplace (1810) showed that the pmf of X can be approximated by the pdf of a certain continuous random variable:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{1}{\sqrt{2\pi np(1-p)}}\, e^{-(k-np)^2/2np(1-p)}.$$

The approximation is best for values of k near the mean $E[X] = np$. This formula can be used to estimate binomial probabilities for large values of n. For example, if $n = 3600$ and $p = 1/2$ then we have

$$P(1770 \leqslant X \leqslant 1830) \approx \int_{1770-0.5}^{1830+0.5} \frac{1}{\sqrt{1800\pi}}\, e^{-(x-1800)^2/1800}\,dx \approx 69.07\%.$$

- **Normal Distributions and the CLT.** More generally, the *normal distribution* with mean μ and σ^2 is defined by the probability density function

$$n(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}.$$

We will write $X \sim N(\mu, \sigma^2)$ to indicate that a continuous random variable X has this pdf. Laplace was the first person to prove a version of the Central Limit Theorem (CLT), which says the following.

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables. For example, these might be random samples taken from some fixed underlying distribution with mean μ and variance σ^2. We define the sample mean $\overline{X} = (X_1 + \cdots + X_n)/n$, and one can check that

$$E[\overline{X}] = \mu \qquad \text{and} \qquad \text{Var}(\overline{X}) = \sigma^2/n.$$

No matter the shape of the underlying distribution, if n is large then Laplace proved that the sample mean is approximately normal:

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} \approx N(\mu, \sigma^2/n).$$

This is the most important general result in all of statistics.

- **The Standard Normal Distribution.** Because of the CLT we want to be able to work with normal distributions. The first important fact is that any normal distribution can be "standardized":

$$X \sim N(\mu, \sigma^2) \qquad \Longleftrightarrow \qquad Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

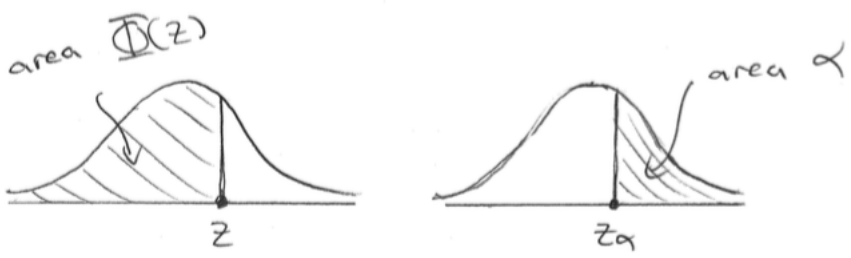
If $Z \sim N(0, 1)$ has a standard normal distribution then we define its *cumulative density function* (cdf) by

$$\Phi(z) = P(Z \leqslant z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$$

The values of $\Phi(z)$ can be looked up in a table. Furthermore, if $0 < \alpha < 1$ is any probability value (i.e., "P-value) then we define the *critical value* z_α to be the unique number with the property

$$\int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = P(Z \geqslant z_\alpha) = \alpha.$$

These numbers can also be looked up in a table. Here are some pictures:



- **Application: Confidence Intervals.** Suppose that we want to estimate an unknown parameter θ of a certain interesting population. To do this we will take a random sample $X_1, X_2, \ldots, X_n$ and then let $\hat{\theta}$ be some estimator that we can compute from the sample. We call this an *unbiased estimator* if $E[\hat{\theta}] = \theta$. To compute a one- or two-sided $(1-\alpha)100\%$ confidence interval we must find real numbers e_α and $e_{\alpha/2}$ with the properties

$$\begin{aligned} P(\hat{\theta} - e_{\alpha/2} < \theta < \hat{\theta} + e_{\alpha/2}) &= 1-\alpha, \\ P(\hat{\theta} - e_\alpha < \theta) &= 1-\alpha, \\ P(\theta < \hat{\theta} + e_\alpha) &= 1-\alpha. \end{aligned}$$

  The interpretation is that the true unknown value of θ will fall inside such a random interval in $(1-\alpha)100\%$ runs of the sampling experiment. The traditional "P-value" is $\alpha = 0.05$.

  If the random variable $\hat{\theta}$ has an approximately **normal** distribution then we can often use the following equations as a starting point for our computations:

$$\begin{aligned} P(-z_{\alpha/2} < Z < z_{\alpha/2}) &= 1-\alpha, \\ P(-z_\alpha < Z) &= 1-\alpha, \\ P(Z < z_\alpha) &= 1-\alpha. \end{aligned}$$

- **Confidence Intervals for a Proportion.** Let X have a Bernoulli distribution with unknown parameter p. In order to estimate p we take a random sample $X_1, \ldots, X_n$ and let $\hat{p} = \overline{X} = (X_1 + \cdots + X_n)/n$. If n is large, then since $E[\overline{X}] = p$ and $\mathrm{Var}(\overline{X}) = p(1-p)/n$, we know from the CLT that

$$\frac{\overline{X} - p}{p(1-p)/n} \approx N(0,1).$$

  Therefore for any P-value $0 < \alpha < 1$ we have

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - p}{p(1-p)/n} < z_{\alpha/2}\right) \approx 1-\alpha,$$

  which after a little algebraic manipulation becomes

$$P\left(\overline{X} - z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} < p < \overline{X} + z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}\right) \approx 1-\alpha.$$

  Unfortunately, the bounds of the confidence interval involve the unknown parameter p. There are several ways to fix this, the easiest of which is to use the crude approximation $p \approx \hat{p}$. This yields the $(1-\alpha)100\%$ confidence interval

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

  which is valid when n is large.

- **Confidence Intervals for the Mean of a Normal Distribution.** Let X have a normal distribution with mean μ and variance σ^2. In order to estimate the unknown μ we take a random sample $X_1, \ldots, X_n$ and let $\hat{\mu} = \overline{X} = (X_1 + \cdots + X_n)/n$. Since $E[\overline{X}] = \mu$ and $\text{Var}(\overline{X}) = \sigma^2/n$, we know from properties of normal distributions that

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

Therefore for any P-value $0 < \alpha < 1$ we have

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha,$$

which after a little algebraic manipulation becomes

$$P\left(\overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

If the standard deviation σ is known to us then we obtain the following exact $(1-\alpha)100\%$ confidence interval for the unknown μ:

$$\overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

However, since μ is unknown to us it is very unlikely that we will know σ, in which case we will estimate σ it with the *sample standard deviation*:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

If n is large then we obtain an approximate $(1-\alpha)100\%$ confidence interval for μ:

$$\overline{X} \pm z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}.$$

If n is small then will use the fact[28] that $(\overline{X} - \mu)/(S/\sqrt{n})$ has a "t distribution with $n-1$ degrees of freedom" to compute an exact $(1-\alpha)100\%$ interval for μ:

$$\overline{X} \pm t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}}.$$

One-sided intervals can be computed in a similar way by replacing $\alpha/2$ with α and using $+$ or $-$ instead of $\pm$.

[28] Unexplained.

Dec 19 (Hurricane Irma Bonus Lecture)

To end this course I will give you a glimpse of *Bayesian statistics*, which is an alternative to the more classical methods that we discussed before the exam. Bayesian techniques are among the newest and the oldest ideas in statistics. Oldest, because these are the methods that were first attempted by Thomas Bayes and Pierre-Simon Laplace in the late 1700s. And newest, because the methods are computationally difficult and are becoming more popular as computers get faster.

In Part I of this course we discussed the notions of *conditional probability* and *Bayes' theorem*. In fact, the reverend Thomas Bayes (1701–1761) never published this result; his notes were edited and published posthumously by Richard Price in 1763 under the title *An Essay towards solving a Problem in the Doctrine of Chances*. We will discuss this work today.

The work begins by clearly stating the problem (in English).

Bayes' Problem (1763).

> *Given* the number of times ion which an unknown event has happende and failed:
> *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.

///

In other words, there exists an unknown "coin" with an unknown probability p of "heads." The coin is "flipped" n times and heads shows up k times. Given this data we are asked to find the probability the probabiliy that p falls in any given interval: $a < p < b$.

Here is a summary of our previous approach to the problem.

Classical Approach to the Problem. We think of p as a fixed constant. The value of p is unknown to us, but it is known to God (or Nature, in the parlance of the 1700s). In order to estimate p we flip the coin n times and let X be the number of heads we get. Then we let $\hat{p} = X/n$ be our estimator for p. If n is large then the Central Limit Theorem says that

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \text{ is approximately a standard normal random variable.}$$

And from this we obtain the confidence interval

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) \approx 1 - \alpha$$

for any value $0 < \alpha < 1$. ///

Unfortunately, we see that the unknown constant p occurs in the upper and lower bounds of the confidence interval. At this point we boldly substituted $\hat{p}$ for p in these bounds. However,

this is a mathematically sloppy thing to do. In addition, we can only use this method to find **symmetric** confidence intervals around p, whereas Bayes' problem wants us to compute $P(a < p < b)$ for any numbers $a < b$.

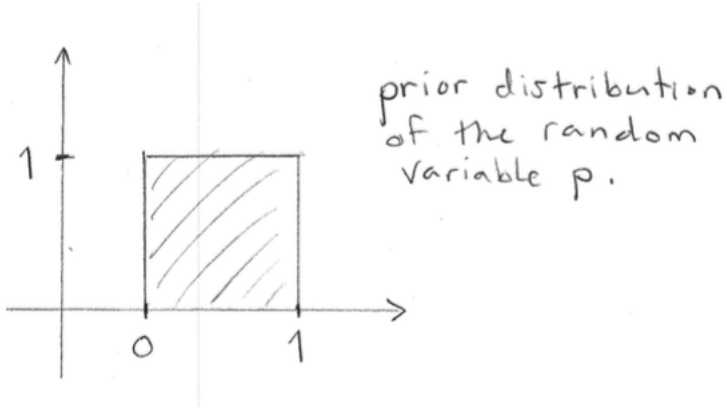
Actually, Bayes' problem is nonsensical from the classical point of view. Since a, b and p are all **constant** (i.e., not random), we either have

$$P(a < p < b) = 1 \qquad \text{or} \qquad P(a < p < b) = 0,$$

depending on whether the statement "$a < p < b$" is true or false.

Here is how Bayes and Price approached the problem.

Bayesian Approach to the Problem. Instead of viewing p as an unknown constant, we will think of p as a **random variable** with a certain density that represents our always imperfect knowledge of p. As we gain information through experiment, the density of p will update to incorporate this new information.[29] Before any experiments were performed, Bayes supposed that all values of p were equally likely. In other words, he assumed that the *prior distribution* of p is uniform on the interval $[0, 1]$:



Thus, to begin with, we have $P(a < p < b) = 1/(b - a)$ for any $0 \leqslant a \leqslant b \leqslant 1$. In order to gain more information about p we perform the following experiment: Flip the coin n times and let X be the number of heads that we get. Note that the random variables X and p are certainly **not independent**. Assume that we perform the experiment and get $X = k$. Then the new distribution of p is given by the *conditional probability*

$$P(a < p < b \,|\, X = k) = \frac{P(a < p < b \text{ and } X = k)}{P(X = k)}.$$

The only issue now is to compute the probabilities on the right. ///

[29] The distinction between classical and Bayesian statistics is akin to the distinction between classical and quantum physics. In the classical picture an electron is a point particle with an actual position that we want to measure. In quantum physics an electron does not have a definite position. Instead is has a "wave function," which encodes its position as a probability distribution.

Before showing you Bayes' solution, I will illustrate the concepts with a toy example.

Toy Example of Bayesian Statistics. Suppose that there are three bowls on a table containing red and white chips, as follows:

The table is locked in a secret room to which only our friend has the key. We send our friend into the room and he comes back with a **red** chip.

Problem: Which bowl did the chip come from?

We could just ask our friend from which bowl he pulled the chip. However, in this scenario he is not allowed to tell us; are we are allowed to know is the color of the chip. Before the experiment is performed we have **no information** about the chip. In this case it is reasonable to assume that all three bowls are equally likely. This will be our *prior distribution*:

i	1	2	3
$P(B_i)$	1/3	1/3	1/3

After we find out that the chip is red, we should update this distribution to reflect the new information. In particular, we should replace the *prior distribution* $P(B_1), P(B_2), P(B_3)$ with the *posterior distribution* $P(B_1|R), P(B_2|R), P(B_3|R)$ where R is the event that "the chip is red." According to the definition of conditional probability we have

$$P(B_i|R) = \frac{P(B_i \cap R)}{P(R)} \qquad \text{and} \qquad P(R|B_i) = \frac{P(R \cap B_i)}{P(B_i)}.$$

Since $P(B_i \cap R) = P(R \cap B_i)$ we can combine these equations to obtain *Bayes' theorem*:

$$P(B_i|R) = \frac{P(B_i)P(R|B_i)}{P(R)}.$$

We can also compute the probability $P(R)$ using the *law of total probability*:

$$\begin{aligned} R &= (R \cap B_1) \sqcup (R \cap B_2) \sqcup (R \cap B_3) \\ P(R) &= P(R \cap B_1) + P(R \cap B_2) + P(R \cap B_3) \end{aligned}$$

$$P(R) = P(B_1)P(R|B_1) + P(B_2)P(R|B_2) + P(B_3)P(R|B_3).$$

In summary we obtain the following formula, which shows us how to obtain the posterior distribution from the prior distribution:

$$P(B_i|R) = \frac{P(B_i)P(R|B_i)}{P(B_1)P(R|B_1) + P(B_2)P(R|B_2) + P(B_3)P(R|B_3)}.$$

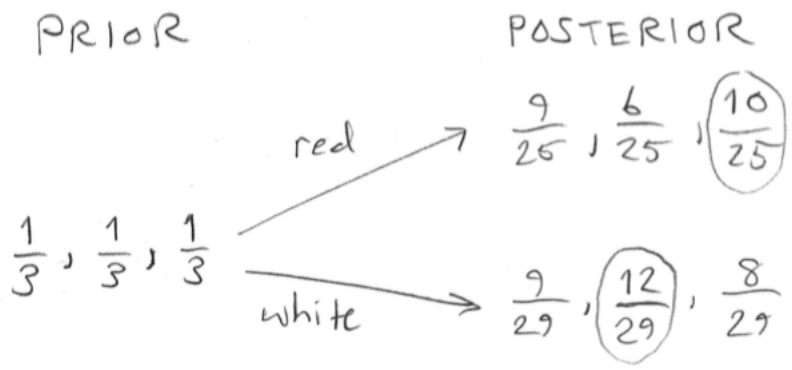
Recall that we have chosen the "uninformative" prior distribution $P(B_1) = P(B_2) = P(B_3) = 1/3$. Since we know the chips in each bowl we can also compute

$$P(R|B_1) = \frac{2}{2+2} = \frac{1}{2}, \quad P(R|B_2) = \frac{1}{1+2} = \frac{1}{3}, \quad P(R|B_3) = \frac{5}{5+4} = \frac{5}{9}.$$

Finally, by plugging in these values we obtain the posterior distribution when the chip is red. For fun, I also calculated the posterior distribution when the chip is white:

i	1	2	3
$P(B_i)$	1/3	1/3	1/3
$P(B_i\|R)$	9/25	6/25	10/25
$P(B_i\|W)$	9/29	12/29	8/29

Here's a picture:

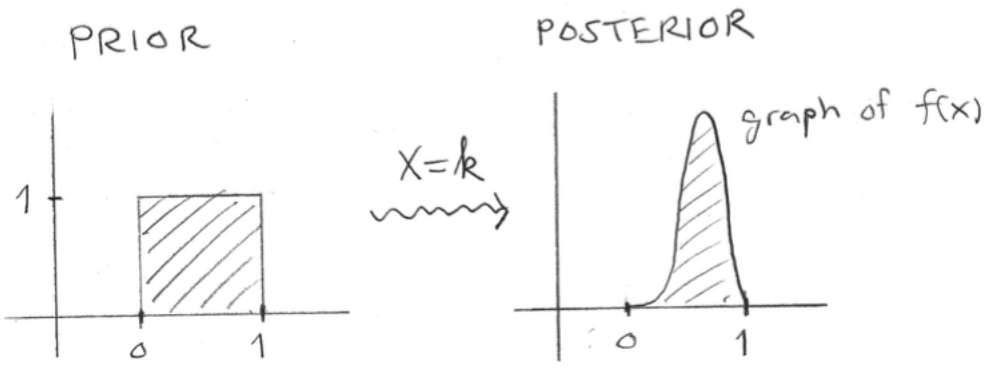


We still don't know which bowl the chip came from, but at least we can now make an educated guess: If the chip is red, it probably came from bowl 3. If the chip is white, it probably came from bowl 2. ///

In the case of Bayes' problem we want to perform the same steps, but with continuous distributions instead of discrete. In other words, we are looking for the density function $f(x)$ with the property

$$P(a < p < b \,|\, X = k) = \int_a^b f(\theta)\, d\theta,$$

which we interpret as follows:



Bayes used a geometric argument to show that

$$f(\theta) = (n+1)\binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

To explain this formula, let's go back to the toy example. Let X, Y be **discrete** random variables and let Θ be a **continuous** random variable. Suppose that $P(Y = \ell)$ is the *prior pmf* of Y. The *posterior pmf* of $Y \,|\, X = k$ is determined by Bayes' theorem

$$P(Y = \ell \,|\, X = k) = \frac{P(Y = \ell) \cdot P(X = k \,|\, Y = \ell)}{P(X = k)},$$

and the denominator $P(X = k)$ can be expanded using the law of total probability:

$$P(X = k) = \sum_{\ell} P(Y = \ell) \cdot P(X = k \,|\, Y = \ell).$$

Now let $f_\Theta(\theta)$ be the *prior pdf* of Θ. Since Θ is continuous, the *posterior pdf* of $\Theta \,|\, X = k$ must be defined by some function $f_{\Theta|X=k}(\theta)$ satisfying

$$P(a < \Theta < b \,|\, X = k) = \int_a^b f_{\Theta|X=k}(\theta)\, d\theta \qquad \text{for all real numbers } a \leqslant b.$$

To compute the function $f_{\Theta|X=k}(\theta)$ we use the same Bayes' theorem, but this time we replace the mass functions $P(Y = \ell)$ and $P(Y = \ell \,|\, X = k)$ by their corresponding densities:

$$f_{\Theta|X=k}(\theta) = \frac{f_\Theta(\theta) \cdot P(X = k \,|\, \Theta = \theta)}{P(X = k)},$$

and we replace the sum in the denominator by the corresponding integral:

$$P(X = k) = \int_{-\infty}^{\infty} f_\Theta(\theta) \cdot P(X = k \,|\, \Theta = \theta).$$

///

Now let's return to Bayes' problem. In this case, X is the number of heads in n flips of the coin and $\Theta = p$ is the underlying probability of heads. Just as with the bowls in the toy example, we will use the "uninformative" prior distribution, i.e., we will assume that all possible values of p are equally likely. In the continuous case this means that we will use a *uniform pdf*:

$$f_p(\theta) = \begin{cases} 1 & \text{when } 0 \leqslant \theta \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, we know for any **specific** value $p = \theta$ that X has a *binomial pmf*:

$$P(X = k \,|\, p = \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

Putting these ingredients together gives us a formula for the posterior density of p, assuming that that we flipped the coin n times and obtained heads $X = k$ times:

$$f_{p|X=k}(\theta) = \frac{f_p(\theta) \cdot P(X = k \,|\, p = \theta)}{\int_{-\infty}^{\infty} f_p(\theta) \cdot P(X = k \,|\, p = \theta)} = \frac{\binom{n}{k} \theta^k (1-\theta)^{n-k}}{\int_0^1 \binom{n}{k} t^k (1-t)^{n-k}\, dt}.$$

One can also show[30] that the integral in the denominator evaluates to

$$\boxed{P(X = k) = \int_0^1 \binom{n}{k} t^k (1-t)^{n-k}\, dt = \frac{1}{n+1}.}$$

This formula is very interesting. It tells us that if the probability of heads $P(H) = p$ has the uniform distribution on $[0, 1]$ then the number of heads X in n flips has the uniform distribution on $k = 0, 1, \ldots, n+1$. This fits very well with the assumption that we know nothing about the coin.

Finally, we can verify that Bayes' formula for the posterior density of $p \,|\, X = k$ is correct:

$$\boxed{f_{p|X=k}(\theta) = (n+1)\binom{n}{k} \theta^k (1-\theta)^{n-k}.}$$

That was a lot of work. Let's get on to solving the problem.

Bayesian Solution to the Problem. Suppose there is a coin with $P(H) = p$ and that this coin is completely unknown to us. To express our lack of knowledge about the coin we will say that p begins as a random variable with a uniform prior distribution:

$$f_p(\theta) = \begin{cases} 1 & \text{when } 0 \leqslant \theta \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

To gain knowledge about the coin we ask a friend to go into the secret room where it is stored, to flip the coin n times and record the number X of heads. If our friend comes back and

[30] omitted

reports that $X = k$ then our new state of knowledge is encoded by the following posterior distribution:

$$f_{p|X=k}(\theta) = (n+1)\binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

Using this knowledge we can compute the probability that the unknown p falls between any real numbers $0 \leqslant a \leqslant b \leqslant 1$:

$$P(a < p < b \,|\, X = k) = \int_a^b (n+1)\binom{n}{k}\theta^k(1-\theta)^{n-k}\,d\theta.$$

///

Small Example. Suppose that an unknown coin is flipped $n = 20$ times and $X = 14$ heads are obtained. Assuming that the probability of heads p has a uniform prior distribution, compute the posterior probability that $p \leqslant 1/2$.

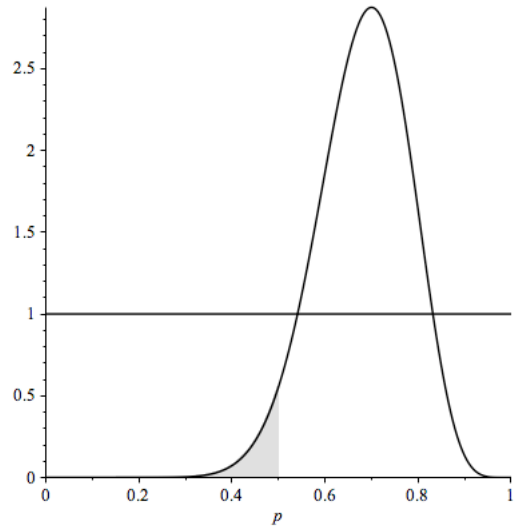
Solution. According to Bayes' formula the posterior density of p is

$$f_{p|X=14}(\theta) = 15\binom{20}{14}\theta^{14}(1-\theta)^{20-14}.$$

Then my laptop computes the probability:

$$P(p \leqslant 1/2 \,|\, X = 14) = \int_0^{1/2} 15\binom{20}{14}\theta^{14}(1-\theta)^{20-14}\,d\theta = 3.92\%.$$

In other words, we can declare with more than 96% probability that this coin favors heads. Here is a picture showing the prior and posterior densities of p, with the shaded area indicating the conditional probability $P(p \leqslant 1/2 \,|\, X = 14)$:



Historical Example. Laplace, births in Paris. Integral is impossible to compute so Laplace used a normal approximation to the posterior. We can get a crude upper bound with Chebyshev's inequality.

Beta distributions. Laplace's rule of succession `https://en.wikipedia.org/wiki/Rule_of_succession`