7. continuous random variables
**Discrete** random variable: takes values in a finite or countable set, e.g.

- $X \in \{1,2,\ldots,6\}$ with equal probability
- $X$ is positive integer $i$ with probability $2^{-i}$

**Continuous** random variable: takes values in an uncountable set, e.g.

- $X$ is the weight of a random person (a real number)
- $X$ is a randomly selected point inside a unit square
- $X$ is the waiting time until the next packet arrives at the server
f(x): \( \mathbb{R} \rightarrow \mathbb{R} \), the *probability density function* (or simply “density”)

Require:

\[
\begin{align*}
  &f(x) \geq 0, \text{ and} \\
  &\int_{-\infty}^{+\infty} f(x) \, dx = 1
\end{align*}
\]

\( \text{i.e., distribution is:} \)

\( \text{\leftarrow \text{ nonnegative, and} \)}

\( \text{\leftarrow \text{ normalized,} \)}

\( \text{just like discrete PMF} \)
F(x): the *cumulative distribution function* (aka the “distribution”)

\[ F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x) \, dx \quad \text{(Area left of a)} \]

\[ P(a < X \leq b) = F(b) - F(a) \quad \text{(Area between a and b)} \]

A key relationship:

\[ f(x) = \frac{d}{dx} F(x), \text{ since } F(a) = \int_{-\infty}^{a} f(x) \, dx, \]
Densities are *not* probabilities; e.g. may be $> 1$

$$P(X = a) = \lim_{\varepsilon \to 0} P(a - \varepsilon < X \leq a) = F(a) - F(a) = 0$$

I.e., the probability that a continuous random variable falls *at* a specified point is *zero*

$$P(a - \varepsilon/2 < X \leq a + \varepsilon/2) =$$

$$F(a + \varepsilon/2) - F(a - \varepsilon/2)$$

$$\approx \varepsilon \cdot f(a)$$

I.e., The probability that it falls *near* that point is proportional to the density; in a large random sample, expect more samples where density is higher (hence the name “density”).
Much of what we did with discrete r.v.s carries over almost unchanged, with $\Sigma_\ldots$ replaced by $\int \ldots \, dx$

E.g.

For discrete r.v. $X$, $E[X] = \Sigma_x xp(x)$

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

Why?

(a) We define it that way

(b) The probability that $X$ falls “near” $x$, say within $x \pm dx/2$, is $\approx f(x)dx$, so the “average” $X$ should be $\approx \Sigma xf(x)dx$ (summed over grid points spaced $dx$ apart on the real line) and the limit of that as $dx \to 0$ is $\int xf(x)dx$
Let \( f(x) = \begin{cases} 1 & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases} \)

\[
F(a) = \int_{-\infty}^{a} f(x) \, dx = \begin{cases} 
0 & \text{if } a \leq 0 \\
 a & \text{if } 0 < a \leq 1 \text{ (since } a = \int_{0}^{a} 1 \, dx) \\
1 & \text{if } 1 < a 
\end{cases}
\]

\[
E[X] = \int_{-\infty}^{\infty} x \, f(x) \, dx = \int_{0}^{1} x \, dx = \frac{x^2}{2} \bigg|_{0}^{1} = \frac{1}{2}
\]

\[
E[X^2] = \int_{-\infty}^{\infty} x^2 \, f(x) \, dx = \int_{0}^{1} x^2 \, dx = \frac{x^3}{3} \bigg|_{0}^{1} = \frac{1}{3}
\]

\[
\text{Var}[X] = E[X^2] - (E[X])^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12} \quad (\sigma \approx 0.29)
\]
properties of expectation

Linearity

\[ E[aX+b] = aE[X]+b \]
\[ E[X+Y] = E[X]+E[Y] \]

Functions of a random variable

\[ E[g(X)] = \int g(x)f(x)dx \]

Alternatively, let \( Y = g(X) \), find the density of \( Y \), say \( f_Y \), (see B&T 4.1; somewhat like r.v. slides 33-35) and directly compute \( E[Y] = \int yf_Y(y)dy \).
Definition is same as in the discrete case

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

Identity still holds:

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

proof “same”
Let \( f(x) = \begin{cases} 
1 & \text{for } 0 < x < 1 \\
0 & \text{elsewhere} 
\end{cases} \)

\[
F(a) = \int_{-\infty}^{a} f(x)dx \\
= \begin{cases} 
0 & \text{if } a \leq 0 \\
1 & \text{if } 1 < a \\
a & \text{if } 0 < a \leq 1 \text{ (since } a = \int_{0}^{a} 1dx) 
\end{cases}
\]

\[
E[X] = \int_{-\infty}^{\infty} x f(x)dx = \int_{0}^{1} x dx = \left. \frac{x^2}{2} \right|_{0}^{1} = \frac{1}{2}
\]

\[
E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_{0}^{1} x^2 dx = \left. \frac{x^3}{3} \right|_{0}^{1} = \frac{1}{3}
\]

Var\[X\] = \(E[X^2] - (E[X])^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12} \quad (\sigma \approx 0.29)\)
Continuous random variable $X$ has density $f(x)$, and

\[ \Pr(a \leq X \leq b) = \int_a^b f(x) \, dx \]

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

\[ E[X^2] = \int_{-\infty}^{\infty} x^2 \cdot f(x) \, dx \]
The Uniform Density Function $\text{Uni}(0.5, 1.0)$

A uniform random variable $X \sim \text{Uni}(\alpha, \beta)$ is uniform in $[\alpha, \beta]$. The probability density function (PDF) is given by:

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & x \in [\alpha, \beta] \\ 0 & \text{otherwise} \end{cases}$$
$X \sim \text{Uni}(\alpha, \beta)$ is uniform in $[\alpha, \beta]$

$$f(x) = \begin{cases} 
\frac{1}{\beta - \alpha} & x \in [\alpha, \beta] \\
0 & \text{otherwise} 
\end{cases}$$

$$\Pr(a \leq X \leq b) = \int_{a}^{b} f(x) \, dx = \frac{b - a}{\beta - \alpha}$$

if $\alpha \leq a \leq b \leq \beta$:

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x) \, dx = \frac{\alpha + \beta}{2}$$

Yes, you should review your basic calculus; e.g., these 2 integrals would be good practice.
uniform random variable: example

\( X \sim \text{Uni}(\alpha, \beta) \) is uniform in \([\alpha, \beta]\)

\[
f(x) = \begin{cases} 
\frac{1}{\beta - \alpha} & x \in [\alpha, \beta] \\
0 & \text{otherwise}
\end{cases}
\]

You want to read a disk sector from a 7200rpm disk drive. Let \( T \) be the time you wait, in milliseconds, after the disk head is positioned over the correct track, until the desired sector rotates under the head.

\( T \sim \text{Uni}(0, 8.33) \)

Average Wait? 4.17ms
waiting for “events”

Radioactive decay: How long until the next alpha particle?

Customers: how long until the next customer/packet arrives at the checkout stand/server?

Buses: How long until the next #71 bus arrives on the Ave?

Yes, they have a schedule, but given the vagaries of traffic, riders with-bikes-and-baby-carriages, etc., can they stick to it?

Assuming events are independent, happening at some fixed average rate of $\lambda$ per unit time – the waiting time until the next event is exponentially distributed (next slide)
$X \sim \text{Exp}(\lambda)$

The Exponential Density Function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$
exponential random variables

\[ X \sim \text{Exp}(\lambda) \]

\[
f(x) = \begin{cases} 
\lambda e^{-\lambda x} & x \geq 0 \\
0 & x < 0 
\end{cases}
\]

\[
E[X] = \frac{1}{\lambda} \quad \text{Var}[X] = \frac{1}{\lambda^2}
\]

\[
\Pr(X \geq t) = e^{-\lambda t} = 1 - F(t)
\]

Memorylessness:

\[
\Pr(X > s + t \mid X > s) = \Pr(X > t)
\]

Assuming exp distr, if you’ve waited \( s \) minutes, prob of waiting \( t \) more is exactly same as \( s = 0 \)
Gambler’s fallacy: “I’m due for a win”

Relation to the Poisson: same process, different measures:

Poisson: *how many* events in a *fixed time*;
Exponential: *how long* until the *next event*

\[ \lambda \text{ is avg # per unit time;} \]
\[ 1/\lambda \text{ is mean wait} \]

Relation to geometric: Geometric is discrete analog:

\[
\text{How long to a Head, 1 flip per sec, prob } p \text{ vs} \\
\text{How long to a Head, 2 flips per sec, prob } p/2, \text{ vs} \\
\text{How long to a Head, 3 flips per sec, prob } p/3, \text{ vs} \\
\vdots \\
\text{Limit is exponential with parameter } 1/p
\]

\[ \text{All have same mean} \]

\[ \text{see also B&T fig 3.8} \]
A brief message from the Math SuperPAC

(This message not approved by any political candidate ...)

interlude
If the election for Governor of Washington were held today, would you vote for Jay Inslee, who prefers the Democratic Party, or Rob McKenna, who prefers the Republican Party?

<table>
<thead>
<tr>
<th></th>
<th>Registered Voters</th>
<th>Likely Voters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inslee – certain</td>
<td>39.5%</td>
<td>41.9%</td>
</tr>
<tr>
<td>Inslee – could change</td>
<td>5.4%</td>
<td>4.4%</td>
</tr>
<tr>
<td>Undecided – lean Inslee</td>
<td>2.3%</td>
<td>2.4%</td>
</tr>
<tr>
<td>Undecided</td>
<td>7.4%</td>
<td>5.8%</td>
</tr>
<tr>
<td>Undecided – lean McKenna</td>
<td>1.8%</td>
<td>1.0%</td>
</tr>
<tr>
<td>McKenna – could change</td>
<td>3.7%</td>
<td>3.6%</td>
</tr>
<tr>
<td>McKenna – certain</td>
<td>40.0%</td>
<td>41.0%</td>
</tr>
<tr>
<td><strong>Total – Inslee</strong></td>
<td><strong>47.2%</strong></td>
<td><strong>48.7%</strong></td>
</tr>
<tr>
<td><strong>Total – McKenna</strong></td>
<td><strong>45.5%</strong></td>
<td><strong>45.6%</strong></td>
</tr>
</tbody>
</table>

632 likely voters: +/- 3.9%


Many registered voters

Suppose a fraction $p$ of them will vote for Inslee

Call 632 of them at random, ask who they like

Suppose 48.7% (308) say “Inslee,” [45.6% (288) McKenna]

*Binomial* random variable, mean $pn$, variance $\sigma^2 = p(1-p)n$

If the gap between $M$ & $I$ is greater than, say, $2\sigma$, we can be reasonably sure the poll difference is “real,” but prediction is sketchy if the gap is smaller. I.e., “margin of error” is $\sim 2\sigma$

PMF for $X_1 \sim \text{Bin}(632, 0.487)$, $X_2 \sim \text{Bin}(632, 0.456)$

\[
\begin{align*}
\sigma^2 &= p(1-p)n = 156.8 \\
\sigma &= \sqrt{p(1-p)n} = 12.5 \\
\sigma/n &= \sqrt{p(1-p)/n} = 1.9%
\end{align*}
\]
X is a normal (aka Gaussian) random variable $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$E[X] = \mu \quad \text{Var}[X] = \sigma^2$$

The Standard Normal Density Function
changing $\mu$, $\sigma$

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Density at $\mu$ is $\approx \frac{.399}{\sigma}$
normal random variables

$X$ is a normal random variable $X \sim N(\mu, \sigma^2)$

$$Y = aX + b$$

$$E[Y] = E[aX+b] = a\mu + b$$

$$\text{Var}[Y] = \text{Var}[aX+b] = a^2\sigma^2$$

$$Y \sim N(a\mu + b, a^2\sigma^2)$$

Important special case: $Z = (X-\mu)/\sigma \sim N(0,1)$

$Z \sim N(0,1)$ “standard (or unit) normal”

Use $\Phi(z)$ to denote CDF, i.e.

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

no closed form 😞
### Table of the Standard Normal Cumulative Distribution Function $\Phi(Z)$

<table>
<thead>
<tr>
<th>$z$</th>
<th>0.00</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.5000</td>
<td>0.5040</td>
<td>0.5080</td>
<td>0.5120</td>
<td>0.5160</td>
<td>0.5199</td>
<td>0.5239</td>
<td>0.5279</td>
<td>0.5319</td>
</tr>
<tr>
<td>0.1</td>
<td>0.5398</td>
<td>0.5438</td>
<td>0.5478</td>
<td>0.5517</td>
<td>0.5557</td>
<td>0.5596</td>
<td>0.5636</td>
<td>0.5675</td>
<td>0.5714</td>
</tr>
<tr>
<td>0.2</td>
<td>0.5793</td>
<td>0.5832</td>
<td>0.5871</td>
<td>0.5910</td>
<td>0.5948</td>
<td>0.5987</td>
<td>0.6026</td>
<td>0.6064</td>
<td>0.6103</td>
</tr>
<tr>
<td>0.3</td>
<td>0.6179</td>
<td>0.6217</td>
<td>0.6255</td>
<td>0.6293</td>
<td>0.6331</td>
<td>0.6368</td>
<td>0.6406</td>
<td>0.6443</td>
<td>0.6480</td>
</tr>
<tr>
<td>0.4</td>
<td>0.6554</td>
<td>0.6591</td>
<td>0.6628</td>
<td>0.6664</td>
<td>0.6700</td>
<td>0.6736</td>
<td>0.6772</td>
<td>0.6808</td>
<td>0.6844</td>
</tr>
<tr>
<td>0.5</td>
<td>0.6915</td>
<td>0.6950</td>
<td>0.6985</td>
<td>0.7019</td>
<td>0.7054</td>
<td>0.7088</td>
<td>0.7123</td>
<td>0.7157</td>
<td>0.7190</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7257</td>
<td>0.7291</td>
<td>0.7324</td>
<td>0.7357</td>
<td>0.7389</td>
<td>0.7422</td>
<td>0.7454</td>
<td>0.7486</td>
<td>0.7517</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7580</td>
<td>0.7611</td>
<td>0.7642</td>
<td>0.7673</td>
<td>0.7704</td>
<td>0.7734</td>
<td>0.7764</td>
<td>0.7794</td>
<td>0.7823</td>
</tr>
<tr>
<td>0.8</td>
<td>0.7881</td>
<td>0.7910</td>
<td>0.7939</td>
<td>0.7967</td>
<td>0.7995</td>
<td>0.8023</td>
<td>0.8051</td>
<td>0.8078</td>
<td>0.8106</td>
</tr>
<tr>
<td>0.9</td>
<td>0.8159</td>
<td>0.8186</td>
<td>0.8212</td>
<td>0.8238</td>
<td>0.8264</td>
<td>0.8289</td>
<td>0.8315</td>
<td>0.8340</td>
<td>0.8365</td>
</tr>
<tr>
<td>1.0</td>
<td>0.8413</td>
<td>0.8438</td>
<td>0.846</td>
<td>0.8485</td>
<td>0.8509</td>
<td>0.8537</td>
<td>0.8565</td>
<td>0.8599</td>
<td>0.8810</td>
</tr>
<tr>
<td>1.1</td>
<td>0.8643</td>
<td>0.8665</td>
<td>0.8680</td>
<td>0.8708</td>
<td>0.8729</td>
<td>0.8750</td>
<td>0.8770</td>
<td>0.8790</td>
<td>0.8980</td>
</tr>
<tr>
<td>1.2</td>
<td>0.8849</td>
<td>0.8869</td>
<td>0.8888</td>
<td>0.8907</td>
<td>0.8925</td>
<td>0.8944</td>
<td>0.8962</td>
<td>0.8980</td>
<td>0.8997</td>
</tr>
<tr>
<td>1.3</td>
<td>0.9032</td>
<td>0.9049</td>
<td>0.9069</td>
<td>0.9082</td>
<td>0.9099</td>
<td>0.9115</td>
<td>0.9131</td>
<td>0.9147</td>
<td>0.9162</td>
</tr>
<tr>
<td>1.4</td>
<td>0.9192</td>
<td>0.9207</td>
<td>0.9221</td>
<td>0.9236</td>
<td>0.9250</td>
<td>0.9264</td>
<td>0.9279</td>
<td>0.9293</td>
<td>0.9306</td>
</tr>
<tr>
<td>1.5</td>
<td>0.9332</td>
<td>0.9345</td>
<td>0.9359</td>
<td>0.9370</td>
<td>0.9382</td>
<td>0.9394</td>
<td>0.9406</td>
<td>0.9418</td>
<td>0.9429</td>
</tr>
<tr>
<td>1.6</td>
<td>0.9452</td>
<td>0.9463</td>
<td>0.9474</td>
<td>0.9484</td>
<td>0.9495</td>
<td>0.9505</td>
<td>0.9515</td>
<td>0.9523</td>
<td>0.9535</td>
</tr>
<tr>
<td>1.7</td>
<td>0.9554</td>
<td>0.9564</td>
<td>0.9573</td>
<td>0.9582</td>
<td>0.9590</td>
<td>0.9598</td>
<td>0.9606</td>
<td>0.9610</td>
<td>0.9625</td>
</tr>
<tr>
<td>1.8</td>
<td>0.9641</td>
<td>0.9649</td>
<td>0.9656</td>
<td>0.9664</td>
<td>0.9671</td>
<td>0.9678</td>
<td>0.9685</td>
<td>0.9690</td>
<td>0.9699</td>
</tr>
<tr>
<td>1.9</td>
<td>0.9713</td>
<td>0.9719</td>
<td>0.9726</td>
<td>0.9732</td>
<td>0.9738</td>
<td>0.9744</td>
<td>0.9750</td>
<td>0.9756</td>
<td>0.9761</td>
</tr>
<tr>
<td>2.0</td>
<td>0.9772</td>
<td>0.9778</td>
<td>0.9783</td>
<td>0.9788</td>
<td>0.9793</td>
<td>0.9798</td>
<td>0.9803</td>
<td>0.9808</td>
<td>0.9812</td>
</tr>
<tr>
<td>2.1</td>
<td>0.9821</td>
<td>0.9826</td>
<td>0.9830</td>
<td>0.9834</td>
<td>0.9838</td>
<td>0.9842</td>
<td>0.9846</td>
<td>0.9850</td>
<td>0.9854</td>
</tr>
<tr>
<td>2.2</td>
<td>0.9861</td>
<td>0.9864</td>
<td>0.9868</td>
<td>0.9871</td>
<td>0.9875</td>
<td>0.9878</td>
<td>0.9881</td>
<td>0.9884</td>
<td>0.9887</td>
</tr>
<tr>
<td>2.3</td>
<td>0.9893</td>
<td>0.9896</td>
<td>0.9898</td>
<td>0.9901</td>
<td>0.9904</td>
<td>0.9906</td>
<td>0.9909</td>
<td>0.9911</td>
<td>0.9913</td>
</tr>
<tr>
<td>3.0</td>
<td>0.9987</td>
<td>0.9987</td>
<td>0.9987</td>
<td>0.9988</td>
<td>0.9988</td>
<td>0.9989</td>
<td>0.9989</td>
<td>0.9989</td>
<td>0.9990</td>
</tr>
<tr>
<td>3.1</td>
<td>0.9990</td>
<td>0.9991</td>
<td>0.9991</td>
<td>0.9991</td>
<td>0.9992</td>
<td>0.9992</td>
<td>0.9992</td>
<td>0.9993</td>
<td>0.9993</td>
</tr>
<tr>
<td>3.2</td>
<td>0.9993</td>
<td>0.9993</td>
<td>0.9994</td>
<td>0.9994</td>
<td>0.9994</td>
<td>0.9994</td>
<td>0.9995</td>
<td>0.9995</td>
<td>0.9995</td>
</tr>
<tr>
<td>3.3</td>
<td>0.9995</td>
<td>0.9995</td>
<td>0.9995</td>
<td>0.9996</td>
<td>0.9996</td>
<td>0.9996</td>
<td>0.9996</td>
<td>0.9996</td>
<td>0.9996</td>
</tr>
<tr>
<td>3.4</td>
<td>0.9997</td>
<td>0.9997</td>
<td>0.9997</td>
<td>0.9997</td>
<td>0.9997</td>
<td>0.9997</td>
<td>0.9997</td>
<td>0.9997</td>
<td>0.9997</td>
</tr>
</tbody>
</table>

**E.g., see B&T p155, p531**

The Standard Normal Density Function

<table>
<thead>
<tr>
<th>$z \sigma$</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>0.1</td>
<td>0.0040</td>
<td>0.0080</td>
<td>0.0120</td>
<td>0.0160</td>
<td>0.0200</td>
<td>0.0240</td>
<td>0.0280</td>
<td>0.0320</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0120</td>
<td>0.0240</td>
<td>0.0360</td>
<td>0.0480</td>
<td>0.0600</td>
<td>0.0720</td>
<td>0.0840</td>
<td>0.0960</td>
</tr>
<tr>
<td>0.3</td>
<td>0.0200</td>
<td>0.0400</td>
<td>0.0600</td>
<td>0.0800</td>
<td>0.1000</td>
<td>0.1200</td>
<td>0.1400</td>
<td>0.1600</td>
</tr>
</tbody>
</table>

### Diagram

- $\mu = 0$
- $\sigma = 1$

E.g., see B&T p155, p531
If $Z \sim N(\mu, \sigma^2)$ what is $P(\mu - \sigma < Z < \mu + \sigma)$?

$P(\mu - \sigma < Z < \mu + \sigma) = \Phi(1) - \Phi(-1) \approx 68\%$

$P(\mu - 2\sigma < Z < \mu + 2\sigma) = \Phi(2) - \Phi(-2) \approx 95\%$

$P(\mu - 3\sigma < Z < \mu + 3\sigma) = \Phi(3) - \Phi(-3) \approx 99\%$

**Why?**

$\mu - k\sigma < Z < \mu + k\sigma \iff -k < (Z - \mu)/\sigma < +k$
**DeMoivre-Laplace Theorem:**

Let $S_n$ = number of successes in $n$ trials (with prob. $p$).

Then, as $n \to \infty$:

$$E[X] = np \quad Var[X] = np(1-p)$$

Equivalently:

**normal approximation to binomial**

$X \sim \text{Bin}(n,p)$ \quad $E[X] = np$ \quad $Var[X] = np(1-p)$

Poisson approx: good for $n$ large, $p$ small ($np$ constant)

Normal approx: For large $n$, ($p$ stays fixed):

$$X \approx Y \sim \mathcal{N}(E[X], Var[X]) = \mathcal{N}(np, np(1-p))$$

Normal approximation good when $np(1-p) \geq 10$

**DeMoivre-Laplace Theorem:**

Let $S_n$ = number of successes in $n$ trials (with prob. $p$).

Then, as $n \to \infty$:

$$Pr\left( a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b \right) \longrightarrow \Phi(b) - \Phi(a)$$

Equivalently:

$$Pr(a \leq S_n \leq b) \longrightarrow \Phi\left( \frac{b-np}{\sqrt{np(1-p)}} \right) - \Phi\left( \frac{a-np}{\sqrt{np(1-p)}} \right)$$
normal approximation to binomial

\[ P(X=k) \]

- Normal(np, np(1-p))
- Binomial(n,p)
- Poisson(np)

\[ n = 100 \]
\[ p = 0.5 \]
DeMoivre-Laplace and the “continuity correction”

Potential pitfalls: Let $S =$ # heads in 100 flips of a fair coin

$$P_r(a \leq S \leq b) \rightarrow \Phi \left( \frac{b-50}{5} \right) - \Phi \left( \frac{a-50}{5} \right)$$

i) $\Pr(50 \leq S \leq 50) \approx .08$, but $\Phi(0) - \Phi(0) = 0$

ii) $\Pr(50.01 \leq S \leq 50.99) = 0$, but $\Phi(.99/5) - \Phi(.01/5) \approx .08$

The “continuity correction”:

Imagine discretizing the normal density by shifting probability mass at non-integer $x$ to the nearest integer (i.e., “rounding” $x$). Then the probability of $S$ falling in the (integer) interval $[a, ..., b]$, inclusive, is $\approx$ the probability of a normal r.v. with the same $\mu, \sigma^2$ falling in the (real) interval $[a-\frac{1}{2}, b+\frac{1}{2}]$.

E.g. i) $\Pr(50 \leq S \leq 50) = \Pr(49.5 \leq S \leq 50.5) \approx \Phi(-0.1) - \Phi(0.1) \approx .08$

ii) $\Pr(50.01 \leq S \leq 50.99) = \Pr($the empty set of integers$) = 0$

normal approximation to binomial

Ex: Fair coin flipped 40 times. Probability of 20 or 21 heads?

Exact answer:

\[ P(X = 20 \lor X = 21) = \left[ \binom{40}{20} + \binom{40}{21} \right] \left( \frac{1}{2} \right)^{40} \approx 0.2448 \]

Normal approximation:

\[ P(20 \leq X < 22) = P(19.5 \leq X \leq 21.5) = P \left( \frac{19.5 - 20}{\sqrt{10}} \leq \frac{X - 20}{\sqrt{10}} \leq \frac{21.5 - 20}{\sqrt{10}} \right) \approx P \left( -0.16 \leq \frac{X - 20}{\sqrt{10}} \leq 0.47 \right) \approx \Phi(0.47) - \Phi(-0.16) \approx 0.2452 \]
Dialog in class:

Q (Student): “Why add/subtract .5? Why not, say, .25?”

A (Prof Evil): “For integer X, the area under the normal density in the strip $X \pm \frac{1}{2}$ is approximately the probability of sampling a normal r.v. that rounds to X, but the area in the strip $X \pm \frac{1}{4}$ is only about half that.”

Q: “What about doubling that area, would that be better?”

A: “Hmm, I dunno, but extrapolating, you could also look at $1/\varepsilon$ times the area in the $X \pm \varepsilon/2$ strip, which in the limit is the density at $X$.”

Graph compares $\pm \frac{1}{2}$ version (green) to density (blue). $\pm \frac{1}{2}$ is better on average, but not uniformly better.
Consider i.i.d. (independent, identically distributed) random vars \( X_1, X_2, X_3, \ldots \)

\( X_i \) has \( \mu = E[X_i] \) and \( \sigma^2 = \text{Var}[X_i] \)

As \( n \to \infty \),

\[
\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma \sqrt{n}} \quad \rightarrow \quad N(0, 1)
\]

Restated: As \( n \to \infty \),

\[
\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N \left( \mu, \frac{\sigma^2}{n} \right)
\]
How tall are you? Why?

Willie Shoemaker & Wilt Chamberlain

Credit: Annie Leibovitz, © 1987?
Human height is approximately normal.

Why might that be true?

R.A. Fisher (1918) noted it would follow from CLT if height were the sum of many independent random effects, e.g. many genetic factors (plus some environmental ones like diet). I.e., suggested part of mechanism by looking at shape of the curve. (WAY before anyone really knew what genes, DNA, etc. were...)

The American Journal of Human Genetics 88, 6–18, January 7, 2011

Table 1. Sixty-Four (and hundreds more probably exist) Loci Showing Significant Evidence for Association with Adult Height, Identified with the Use of the IBC Array

<table>
<thead>
<tr>
<th>Locus Rank</th>
<th>Chr.</th>
<th>Candidate Gene</th>
<th>SNP</th>
<th>Effect Allele</th>
<th>MAF</th>
<th>Effect (cm/allele)</th>
<th>European Ancestry Phase 1 (up to 53,394)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7q22</td>
<td>CDK6</td>
<td>rs4272</td>
<td>A</td>
<td>0.21</td>
<td>−0.46</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6p21</td>
<td>HMGAIN</td>
<td>rs1120781</td>
<td>A</td>
<td>0.00</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>12q15</td>
<td>HMGAIN</td>
<td>rs900157</td>
<td>A</td>
<td>0.17</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>20q11</td>
<td>MMP24</td>
<td>rs1292711</td>
<td>A</td>
<td>0.35</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>17q23</td>
<td>MAP3K3</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>17q24</td>
<td>GH1-GH2</td>
<td>rs380613</td>
<td>A</td>
<td>0.35</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1p36</td>
<td>MFAP2</td>
<td>rs1292711</td>
<td>A</td>
<td>0.35</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>15q26</td>
<td>IGFI1R</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>7p22</td>
<td>GNA12</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>17q23</td>
<td>TBX2</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>12q22</td>
<td>SOCS2</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>9q22</td>
<td>PTCH1</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>14q11</td>
<td>NFATC4</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>15q26</td>
<td>ACAN</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2q24</td>
<td>NPPC</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>6p21</td>
<td>PPARD</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>20q11</td>
<td>MYH7B</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>19q13</td>
<td>IL11</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Continued

<table>
<thead>
<tr>
<th>Locus Rank</th>
<th>Chr.</th>
<th>Candidate Gene</th>
<th>SNP</th>
<th>Effect Allele</th>
<th>MAF</th>
<th>Effect (cm/allele)</th>
<th>European Ancestry Phase 1 (up to 53,394)</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>2p23</td>
<td>LCR</td>
<td>rs780094</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>1q41</td>
<td>TGFB2</td>
<td>rs900157</td>
<td>A</td>
<td>0.17</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>20q11</td>
<td>CDK5RAP1</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>2p12</td>
<td>EII2AK3</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>19p13</td>
<td>INSRA</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>6q25</td>
<td>ESI1</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>6q37</td>
<td>D2S1222</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>6q36</td>
<td>PLC4</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>6q36</td>
<td>RS6K1A</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>15q21</td>
<td>CYP19A1</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>5q31</td>
<td>SLC22A5</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>7q15</td>
<td>JAZF1</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>17q13</td>
<td>POLR2A</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>1p22</td>
<td>PKN2</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>7q22</td>
<td>CNOT4</td>
<td>rs380608</td>
<td>A</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
</tr>
</tbody>
</table>
in the real world…
in the real world…
in the real world…

Histogram of Daily Trading-Related Revenue* — Twelve Months Ended December 31, 2007

*Excludes daily profits and losses in the ABS CDO market, including recent subprime-related losses.
in the real world…
pdf and cdf

\[ f(x) = \frac{d}{dx} F(x) \quad F(a) = \int_{-\infty}^{a} f(x) \, dx \]

sums become integrals, e.g.

\[ E[X] = \sum x \cdot p(x) \quad E[X] = \int_{-\infty}^{\infty} x \cdot f(x) \, dx \]

most familiar properties still hold, e.g.

\[ E[aX+bY+c] = aE[X]+bE[Y]+c \]
\[ Var[X] = E[X^2] - (E[X])^2 \]
Three important examples

**X ~ Uni(α, β) uniform in [α, β]**

\[
f(x) = \begin{cases} 
\frac{1}{\beta-\alpha} & x \in [\alpha, \beta] \\
0 & \text{otherwise}
\end{cases}
\]

\[E[X] = (\alpha+\beta)/2\]
\[\text{Var}[X] = (\alpha-\beta)^2/12\]

**X ~ Exp(λ) exponential**

\[
f(x) = \begin{cases} 
\lambda e^{-\lambda x} & x \geq 0 \\
0 & x < 0
\end{cases}
\]

\[E[X] = \frac{1}{\lambda}\]
\[\text{Var}[X] = \frac{1}{\lambda^2}\]

**X ~ N(μ, σ²) normal (aka Gaussian, aka the big Kahuna)**

\[
f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{- (x-\mu)^2 / 2\sigma^2}
\]

\[E[X] = \mu\]
\[\text{Var}[X] = \sigma^2\]