Lecture 5 (Ch. 1)

Last time: (Sample vs. pop) \( \iff \) (hist. vs. dist.)

Named distributions: \( f(x) \) and \( p(x) \).

1) \( \text{Exp} (\lambda) \)
   \[ f(x) = \lambda e^{-\lambda x} \quad x \geq 0, \quad \lambda > 0 \]

2) \( \text{Pois} (\lambda) \)
   \[ p(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 0, 1, 2, \ldots \infty, \quad \lambda > 0 \]

3) \( \text{Binomial (revisited)} \)
   \( x = \text{discrete} \)
   We'll derive its mass function, next time, but it's:
   - E.g., \# of heads out of \( n \) tosses.
   - \# of defective gates on a chip with \( n \) gates.
   - \# of girls in a sample of size \( n \).

   \( \text{# of H's out of n tosses} \)
   \[ p(X) = \binom{n}{x} \left( \frac{\theta}{2} \right)^x \left( 1 - \frac{\theta}{2} \right)^{n-x}, \quad x = 0, 1, \ldots, n \]
   \( \text{prob of } X \text{ heads out of } n \text{ tosses.} \)

   Parameters: \( n, \theta \). \[ n=\text{integers}, \quad 0 < \theta < 1 \]
   \( \text{Binom} (n, \theta) \).

Depending on the value of the parameters, it can look like:

\[ \text{p(x)} \quad \text{A} \quad \text{gain, note:} \]
\[ \begin{array}{c}
\text{look like} \\
\text{hist, but} \\
\text{no data.}
\end{array} \]

\[ \text{p(x)} \quad \text{In lab, you'll see how the} \]
\[ \text{shape depends} \\
\text{on the} \\
\text{par}am., \]

\[ \text{Small } n \]

\[ \text{large } n \]
4) Normal/Gaussian, \( x = \text{const.} \)

E.g. temperature, height, weight, blood pressure, ...  

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

Note: 
- if \( \mu = 0, \sigma = 1 \)
- Then \( f(x) \) is std. Normal.

\( \mu \) is the measure of location or middle, or centrality.

Parameters/meaning: \( \mu, \sigma \). 
- \( \mu \): measure of spread
- \( \sigma \): inflection point

Note: Symmetric about \( x = \mu \).

\[ N(\mu, \sigma) \quad [\text{some books: } N(\mu, \sigma^2)] \]

Important: Resist the temptation to call \( \mu \) and \( \sigma \) mean and standard deviation, at least for now. Otherwise you'll get very confused. They are simply parameters of the distribution.
So far, we have

\( P(x) = \text{(probability)} \) mass function (pmf) \( \uparrow \) \text{discrete/categ.}

\( f(x) = \text{(probability)} \) density function (pdf) \( \uparrow \) \text{continuous.}

Recall that there is a connection between dists and probs.

\( \Rightarrow \) For discrete/categorical variables: \( \text{prob} = \sum_{x} P(x) \)

For example, if \( X \sim \text{Binom}(n, \pi) \), then

\[
\text{prob} \left( a \leq X \leq b \right) = \sum_{x=a}^{b} \binom{n}{x} \pi^x (1-\pi)^{n-x} P(x)
\]

\( \text{Even though Table II gives some binomial areas, you do NOT have to use it. You can use this formula.} \)

\( \Rightarrow \) For continuous variables: \( \sqrt{\text{prob}} = \int f(x) \, dx = \text{area} \)

For example, if \( X \sim \text{N}(\mu, \sigma) \), then

\[
\text{prob} \left( a < X < b \right) = \int_{a}^{b} \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2} \, dx
\]
However, for the Normal dist'v. it turns out that it is impossible to compute areas (integrals) analytically/except.

Let's start with the simpler std. Norm., i.e. \( N(\mu=0, \sigma=1) \):

\[
\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} \, dx \quad \text{Note: } N(0,1)
\]

Unfortunately, integrals of this type can be done only numerically. Their values are tabulated in Table I.

In the table, \( \int_{-\infty}^{0.23} f(x) \, dx = 0.5910 \) gives 

**Table I**

<table>
<thead>
<tr>
<th>( z^* )</th>
<th>.00</th>
<th>.01</th>
<th>.02</th>
<th>.03</th>
<th>.04</th>
<th>.05</th>
<th>.06</th>
<th>.07</th>
<th>.08</th>
<th>.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>.5000</td>
<td>.5040</td>
<td>.5080</td>
<td>.5120</td>
<td>.5160</td>
<td>.5199</td>
<td>.5239</td>
<td>.5279</td>
<td>.5319</td>
<td>.5359</td>
</tr>
<tr>
<td>0.1</td>
<td>.5398</td>
<td>.5438</td>
<td>.5478</td>
<td>.5517</td>
<td>.5557</td>
<td>.5596</td>
<td>.5636</td>
<td>.5675</td>
<td>.5714</td>
<td>.5753</td>
</tr>
<tr>
<td>0.2</td>
<td>.5793</td>
<td>.5832</td>
<td>.5871</td>
<td>.5910</td>
<td>.5948</td>
<td>.5987</td>
<td>.6026</td>
<td>.6064</td>
<td>.6103</td>
<td>.6141</td>
</tr>
<tr>
<td>0.3</td>
<td>.6179</td>
<td>.6217</td>
<td>.6255</td>
<td>.6293</td>
<td>.6331</td>
<td>.6368</td>
<td>.6406</td>
<td>.6443</td>
<td>.6480</td>
<td>.6517</td>
</tr>
<tr>
<td>0.4</td>
<td>.6554</td>
<td>.6591</td>
<td>.6628</td>
<td>.6664</td>
<td>.6700</td>
<td>.6736</td>
<td>.6772</td>
<td>.6808</td>
<td>.6844</td>
<td>.6879</td>
</tr>
<tr>
<td>0.5</td>
<td>.6915</td>
<td>.6950</td>
<td>.6985</td>
<td>.7019</td>
<td>.7054</td>
<td>.7088</td>
<td>.7123</td>
<td>.7157</td>
<td>.7190</td>
<td>.7224</td>
</tr>
<tr>
<td>0.6</td>
<td>.7257</td>
<td>.7291</td>
<td>.7324</td>
<td>.7357</td>
<td>.7389</td>
<td>.7422</td>
<td>.7454</td>
<td>.7486</td>
<td>.7517</td>
<td>.7549</td>
</tr>
<tr>
<td>0.7</td>
<td>.7580</td>
<td>.7611</td>
<td>.7642</td>
<td>.7673</td>
<td>.7704</td>
<td>.7734</td>
<td>.7764</td>
<td>.7794</td>
<td>.7823</td>
<td>.7852</td>
</tr>
</tbody>
</table>

In 340, use Table I, unless the problem says “By R.”
To find the area between 2 x's, there is a trick:

\[
\text{prob}(a < x < b) = \text{area between } a \text{ and } b =
\]

\[
= \text{prob}(x < b) - \text{prob}(x < a)
\]

Both of these can be obtained from Table I.

**Example:** What is the area under the std. Normal between -1 and +1?

\[
= 0.8413 - 0.1587 = 0.6826
\]

(famous 68% from High School)

**Example:** How about between -2.1 and 0?

\[
= 0.5 - 0.0179 = 0.4821
\]

**Example:** What is the area to the right of (-2.1)?

\[
1 - (0.0179) = 0.9821
\]

or \[
0.4821 + 0.5 = 0.9821
\]
what about \( \mathcal{N}(\mu, \sigma) \)?

It would be impractical to have tables for all \( \mu \) and \( \sigma \)!

Fortunately, there is a trick: \textit{change variables!} also called \textit{standardization}

\[
x \rightarrow z = \frac{x - \mu}{\sigma} \quad (z\text{-score})
\]

So, to compute area between 2 values:

\[
\text{prob}(a < x < b) = \text{prob}
\left( \frac{a - \mu}{\sigma} < \frac{x - \mu}{\sigma} < \frac{b - \mu}{\sigma} \right)
\]

\[
= \text{prob}(z < \frac{b - \mu}{\sigma}) - \text{prob}(z < \frac{a - \mu}{\sigma})
\]

Either way (algebraically or graphically) you can obtain the value of each term from Table 1.

\textbf{Example:} What’s the area between -2 and 2 for a normal curve with \( \mu = 4 \), \( \sigma = 3 \)

\[
\text{prob}(z < \frac{-2 - 4}{3}) - \text{prob}(z < \frac{2 - 4}{3}) = 0.2514 - 0.0228 = 0.2286
\]
Let's turn things around: Given \( f(x) \), and area, find \( x \).

E.g., median: \( \int_{-\infty}^{x} f(x) \, dx = \frac{1}{2} \)

Median is a special case of a more general concept:

\( \frac{n}{100} \)th percentile \( \Rightarrow \int_{-\infty}^{x} f(x) \, dx = \frac{n}{100} \)

\( \text{median} = 50 \text{th percentile} = 0.5 \text{ quantile} = 2^{nd} \text{ quartile} \)

**Example:** What's the 90th percentile (0.9 quantile) of \( N(\mu, \sigma) \)?

**Graphically:**

Algebraically:

\[ 0.9 = \Pr(x < ?) = \Pr\left( \frac{x - \mu}{\sigma} < \frac{? - \mu}{\sigma} \right) = \Pr(z < \frac{? - \mu}{\sigma}) \]

\[ \text{Table I} \rightarrow \frac{? - \mu}{\sigma} = 1.285 \]

\[ ? = \mu + 1.285 \sigma \]

Note: percentile is a number on the \( x \)-axis, not a percent. I.e., a percentile of \( x \) has the same units as \( x \).
Summary

By now, you should be able to (for hists & Dists)

1) compute the area to the left (or right) of $x = a$,
2) compute ... between $x = a$, $x = b$,
3) compute $x = a$, given the area to left (or right).

If the left area is $n\%$, then $x = a$ is called the $n^{th}$ percentile.

If $f(x) = \text{std. normal}$, then Table I.

\[ x \sim \mathcal{N}(0, 1) \]

If $f(x) = \text{normal}(\mu, \sigma)$, then \textbf{standardize first}, and proceed ...

\[ x \sim \mathcal{N}(\mu, \sigma) \]

\[ z = \frac{x - \mu}{\sigma} \]

hw lead 5-1

What's the prob of getting 1 or 2 boys in a sample of size 10, from a population in which the proportion of boys is exactly 50%?

hw lead 5-2

Suppose the density function for $x$ is given by the normal distr. with parameters $\mu, \sigma$. I.e.

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

a) Find the density function ($f(z)$) for $z = \frac{x - \mu}{\sigma}$.

\[ \text{Hint: Start with } \int f(x) \, dx = 1, \text{ do not do the integral, important} \]

but instead, do a change of variables until you get

\[ \int [\ldots] \, dz = 1. \text{ Then } [\ldots] \text{ is } f(z). \]

b) In $f(z)$, in the place where you would expect to find $\mu$ and $\sigma$, what numbers do you see?
If \( x \) follows \( N(\mu, \sigma) \), what's the prob. of \( x \) being within 1 \( \sigma \) of \( \mu \)?

Standardization is important in finding probs. Although it almost always refers to the change of variable \( z = \frac{x - \mu}{\sigma} \), taking \( N(\mu, \sigma) \) to \( N(0,1) \), sometimes a different change of variable is required to obtain something that has \( N(0,1) \) dist.

Find the \( P(x < 2) \) if \( \log_e(x) \) has a std. normal dist.

What's the \( n^{th} \) percentile of the unif distribution over \([a, b] \)?

Hint: Answer will depend on \( a \) and \( b \).

Optional:

a) Use the binomial mass function to show that the prob. of getting “at least 1 head out of \( n \) tosses” is \( 1 - (1 - \pi)^n \), where \( \pi \) is the prob. of getting a head on a single toss. Show work!

b) For \( \pi < 0.5 \), what's the numerical value of that prob, as \( n \to \infty \)?

c) Now, what's the prob. of exactly 1 head out of \( n \) tosses?

d) And what does that prob. converge to, as \( n \to \infty \)?

Hint: You may need L'Hopital's Rule.