NORMAL (GAUSSIAN) DENSITY FAMILY (“Bell shaped curve”)

Most common approximate description of distribution of data)
– Completely specified by mean, $\mu$, & St.Dev., $\sigma$
– Bell shaped, symmetric, unimodal.

Why Normal density?
Justification often comes from Central Limit Theorem.
If $X$ is actually sum of many independent random variables, then it is (close to) normally distributed.
(remember sum of 4 uniform random variables last time)
E.g. often used as a model for heights, IQs, velocity distributions. . . .

Normal Distribution $N(\mu, \sigma^2)$:
The density of $Y$ is
$$f(y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$
The two parameters that are needed to define a normal are: $\mu = E[Y]$, $\sigma^2 = var(Y)$.

Standard Normal Distribution $N(0,1)$: $\mu=0$, $\sigma=1$
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Lemma: If $Z \sim N(0,1)$, and $X = \mu + \sigma Z$ then $X \sim N(\mu, \sigma^2)$
Hence Standardized Z-score $\zeta = \frac{Z-\mu}{\sigma}$=no. of SD’s away from mean.

Mean and Variance: Lemma: If $Z \sim N(0,1)$, then $EZ=0$ and $var Z=1$. 
Recap

Cumulative Distribution Function \( F(y) = P(X \leq y) \)

Quantile  Quantiles of the distribution are points \( z_c \), such that \( P(X \leq z_c) = F(z_c) = c \),

E.g. For a standard normal, the tenth percentile is \( \Phi^{-1}(0.1) = -1.28 \).

What is the 90th percentile?

For Discrete Random Variable: smallest value \( z_c \) so that \( P(X = z) \geq c \)

\[
> \text{pbinom}(0:5, \text{size}=5, p=.1) \\
[1] 0.59049 0.91854 0.99144 0.99954 0.99999 1.00000 \\
> \text{qbinom}(0.6, \text{size}=5, \text{prob}=0.1) \\
[1] 1
\]

Different Distributions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Expected value</th>
<th>Variance</th>
<th>Probability Mass</th>
<th>Density Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial(n,p)</td>
<td>( np )</td>
<td>( np(1-p) )</td>
<td>( P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} )</td>
<td>( \frac{n^k}{k!} )</td>
</tr>
<tr>
<td>Poisson(( \lambda ))</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
<td>( P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} )</td>
<td>( e^{-\lambda} \frac{\lambda^k}{k!} )</td>
</tr>
<tr>
<td>Uniform(a,b)</td>
<td>( \frac{a+b}{2} )</td>
<td>( \frac{(b-a)^2}{12} )</td>
<td>( P(x) = \frac{1}{b-a} )</td>
<td>( \frac{1}{\sqrt{2\pi}\sigma} )</td>
</tr>
<tr>
<td>Normal(( \mu, \sigma^2 ))</td>
<td>( \mu )</td>
<td>( \sigma^2 )</td>
<td>( P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} )</td>
<td>( \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} )</td>
</tr>
</tbody>
</table>

Areas under the N(0,1) curve:

\[
\Phi(z) = \int_{-\infty}^{z} \phi(x) \, dx, \quad \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}
\]

A frequent task is to use the normal tables to go from z-scores to percents and vice-versa. It is a mechanical process, and you should be on top of it.
Illustrate both directions by example:
Suppose men’s heights (in inches) are approximated by $X \sim \mathcal{N}(69, 3^2)$

**A. Scores $\rightarrow$ Percents**

% men between 5’ 4” and 6’ ??
1. State problem: $P\{64 \leq Y \leq 72\}$
2. Standardize:
   $$\frac{64-69}{3} \leq \frac{Y-\mu}{\sigma} \leq \frac{72-69}{3}$$
   $$= P(-1.66 \leq Z \leq 1.0)$$
3. Use table/software to evaluate this probability. (No closed form formula) Using R:
   $$\text{pnorm}(1) - \text{pnorm}(-1.6667) \quad \text{returns} \quad 0.7935577$$

**B. Percents $\rightarrow$ Quantiles** (inverse to A.)
Where is upper decile (top 10% point) of men’s heights?

1. Visualize problem with standard normal
2. Tables $z = 1.28$
   > qnorm(0.9)
   [1] 1.281552
   > qnorm(0.9, 69, 3)
   [1] 72.84465
3. Unstandardize:
   $$x = \mu + \sigma \cdot Z = 69 + 3 \cdot 1.28 = 72.8$$
What does it mean to be 2 or 3 SD’s away from the average??

**68-95-99.7 rule**

- 68% of data lie in $[\mu-\sigma, \mu+\sigma]$ (32 out of 100 don’t)
- 95% of data in $[\mu-2\sigma, \mu+2\sigma]$ (5 in 100 don’t)
- 99.7% of data in $[\mu-3\sigma, \mu+3\sigma]$ (3 in 1000 don’t)

**Unusual ≠ Rare:** E.g. women’s heights $\mu=64.5”$, $\sigma = 2.5”$ $\mu+3\sigma = 72” = 6’$

# women in U.S. $\geq 6’ \approx 270$ mill $\times 1/2 \times (1.5/1000) \approx 200,000$

(even if unusual, you’ll have lots of company!)

**Normal quantile plots**

- *informal, graphical* method
- to assess if data are *nearly* normal
- more sensitive than histograms / stem plots
- more straightforward on a computer (no bin width issues)

Algorithm:

1. Sort data, increasing order, we use the notation $x_{(1)} = min(x_i), x_{(k)} = k$th value after the sample has been ordered.

2. Create corresponding z-scores:
   Roughly: $z_{(i)}$ is where we would expect the (standardized) $x_{(i)}$ to be if the $x_i$ were approximately normal.

This is done by computing $i - 0.5/n$ and then computing the $z_i$ corresponding to the ith ordered Normal value.

We use the 0.5 to make sure the last value doesn’t become infinite.
3. Plot the expected z-scores against the observed values.

   Note: there is no fixed conventions about certain points:
   
   (a) about which axis the normal quantiles are placed
   (b) whether to account for ties

If I were to do it manually in R:

```r
> sort(bwt)
[1] 1150 1173 1250 1250 1300 1450 1650 1700 1700 1700 1750 1800 1950
....................
[85] 3200 3250 3300 3300 3300 3400 3400 3500 3500 3500 3550 3550 3600
[99] 3600 3650 3680 3800 4000 4000 4100 4200 4850
> (((1:107)-0.5)/107)
[1] 0.004672897 0.014018692 0.023364486 0.032710280 0.042056075 0.051401869
..........................
[103] 0.957943925 0.967289720 0.976635514 0.985981308 0.995327103
> qnorm(((1:107)-0.5)/107)
[1] -2.59913892 -2.19676291 -1.98874848 -1.84237333 -1.72730919 -1.63141046
.......[49] -0.11740130 -0.09384329 -0.07033725 -0.04687002 -0.02342858 0.00000000
[55] 0.02342858 0.04687002 0.07033725 0.09384329 0.11740130 0.14102465
.......[103] 1.72730919 1.84237333 1.98874848 2.19676291 2.59913892
> plot(qnorm(((1:107)-0.5)/107),sort(bwt))

OR

> qqnorm(bwt)
> qqnorm(rnorm(107,mean(bwt),sqrt(var(bwt))))
> qqline(rnorm(107,mean(bwt),sqrt(var(bwt))))
```
When the data are on the y-axis then:

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Shape of QQ Plot(^a)</th>
<th>Position of Sample quantiles wrt Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>(near) Normal</td>
<td>Linear</td>
<td>Roughly the same</td>
</tr>
<tr>
<td>Right skew</td>
<td>Elbow up</td>
<td>High quantiles of X are larger than Normal</td>
</tr>
<tr>
<td>Left skew</td>
<td>Elbow down</td>
<td>Low quantiles of X are smaller than Normal</td>
</tr>
<tr>
<td>Heavy Tailed</td>
<td>Backward S</td>
<td>Quantiles are more stretched out – High quantiles of X are larger and Low Quantiles are smaller</td>
</tr>
<tr>
<td>Light Tailed</td>
<td>S</td>
<td>Quantiles are less stretched out – High quantiles of X are smaller and Low Quantiles are higher</td>
</tr>
</tbody>
</table>

\(^a\)When the data are on the x-axis conventions are swapped!

Stragglers < – > outliers Don’t worry about minor wiggles.
Histogram of Left Skewed

Histogram of Heavy Tails

Left Skewed

Heavy Tails
Switched Axes — What is it??

R commands used here

\[
\text{dnorm}(x, \text{mean}=0, \text{sd}=1, \log = \text{FALSE})
\]
\[
\text{pnorm}(q, \text{mean}=0, \text{sd}=1, \text{lower.tail} = \text{TRUE}, \log.p = \text{FALSE})
\]
\[
\text{qnorm}(p, \text{mean}=0, \text{sd}=1, \text{lower.tail} = \text{TRUE}, \log.p = \text{FALSE})
\]
\[
\text{rnorm}(n, \text{mean}=0, \text{sd}=1)
\]

#Plots (xyz a sequence of numbers to plot)
> plot(xyz,\text{dnorm}(xyz,45,5),\text{type}='l',\text{ylim}=c(0,0.15),\text{ylab}='')
> plot(xyz,\text{dnorm}(xyz,30,10),\text{type}='l',\text{ylim}=c(0,0.15),\text{ylab}='')
> plot(xyz,\text{dnorm}(xyz,69,3),\text{type}='l',\text{ylim}=c(0,0.15),\text{ylab}='')

#Between values
> plot(x,\text{dnorm}(x,69,3),\text{type}='l')
> lines(c(64,64),c(0,\text{dnorm}(64,69,3)))
> lines(c(72,72),c(0,\text{dnorm}(72,69,3)))

#Quantiles
> \text{qnorm}(0.1)
> qqnorm(bwt)
> qqline(bwt)