Joint distributions I

Joint distribution of N random variables $X_1, \dots, X_N$:

- Discrete: $f_{\mathbf{X}}(x_1, \dots, x_N) \geq 0, x_i \in R_{X_i}$, for $i = 1, \dots, N$.

$$\sum_{R_1, \dots, R_N} f_{\mathbf{X}}(x_1, \dots, x_N) = 1.$$

- Continuous: $f_{\mathbf{X}}(x_1, \dots, x_N) \geq 0, x_i \in \mathcal{R}, i = 1, \dots, N$.

$$\int_{\mathcal{R}^N} f_{\mathbf{X}}(x_1, \dots, x_N) dx_1 \cdot dx_N = 1.$$

Joint distributions II

Marginals:

- Discrete example:
  $f_{X_2,X_4}(x_2^*, x_4^*) = \sum_{x_1 \in R_{X_1}, x_3 \in R_{X_3}, x_5 \in R_{X_5} \ldots x_N \in R_{X_N}} f_{\mathbf{X}}(x_1, x_2^*, x_3, x_4^*, x_5, \ldots, x_N)$.
- Continuous example: $f_{X_2,X_4}(x_2^*, x_4^*) = \int_{x_1 \in \mathcal{R}, x_3 \in \mathcal{R}, x_5 \in \mathcal{R} \ldots x_N \in \mathcal{R}} f_{\mathbf{X}}(x_1, x_2^*, x_3, x_4^*, x_5, \ldots, x_N) dx_1 dx_3 dx_5 \cdots dx_N$.

Joint distributions III

Conditionals:

- Discrete example
  $f_{X_2|X_4}(x_2|x_4) = \frac{f_{X_2,X_4}(x_2,x_4)}{f_{X_4}(x_4)}$.
  Discrete distribution on R_{X_2} for each value of x_4.

- Continuous example:
  $f_{X_2|X_4}(x_2|x_4) = \frac{f_{X_2,X_4}(x_2,x_4)}{f_{X_4}(x_4)}$.
  Continuous density on $\mathcal{R}$ for each value of x_4.

Mutual independence of N variables:
$f_{\mathbf{X}}(x_1, \ldots, x_N) = \prod_{i=1}^{N} f_{X_i}(x_i)$.
For all values $x_1, \ldots, x_N$.

Covariance of two random variables I

Covariance of two random variables X, Y with means μ_X, μ_Y.
Take $g(x, y) = (x - \mu_X) \cdot (y - \mu_Y)$.
$Cov(X, Y) = Eg(X, Y)$ - A measure of how the variables 'covary'.

$Cov(X, Y) > 0 \longrightarrow$ when X increases Y tends to increase.
$Cov(X, Y) < 0 \longrightarrow$ when X increases Y tends to decrease.

Show that $Cov(aX, bY) = a \cdot bCov(X, Y)$.
Changing the units of a measurement will change covariance.

Correlation $\rho(X, Y)$ does not depend on units of measurement:
$\rho(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$
Show that: $\rho(aX, bY) = \rho(X, Y)$.

Covariance of two random variables II

Independence:

Show that $Cov(X, Y) = EXY - \mu_X \mu_Y$.
Conclude: If X, Y are independent $Cov(X, Y) = 0$. (Converse is not true.)

Variance of a sum of random variables I

$$
\begin{aligned}
Var(X+Y) =& E[(X+Y)^2] - [E(X+Y)]^2 \\
=& E[X^2 + 2XY + Y^2] - [E(X)]^2 - 2E(X)E(Y) - [E(Y)]^2 \\
=& E(X^2) - [E(X)]^2 + E(Y^2) - [E(Y)]^2 \\
& + 2E(XY) - 2E(X)E(Y) \\
=& Var(X) + Var(Y) + 2Cov(X,Y)
\end{aligned}
$$

Conclude: If X, Y independent: $Var(X+Y) = Var(X) + Var(Y)$.

More generally, if $X_1, \ldots, X_N$ are independent then

$Var(\sum_{n=1}^{N} X_n) = \sum_{n=1}^{N} Var(X_n)$.

Properties of sample average I

We draw with replacement form a box N times and record the number on each draw as $X_1, X_2, \ldots, X_N$. Because we draw with replacement we can *assume* that the variables X_n are independent. Let μ_B be the *average of the box* and let σ_B^2 be the mean square deviation (MSD) of the box.

Recall that $EX_n = \mu_B$ and $VarX_n = \sigma_B^2$ for each n.
Denote the sample average as $\overline{X} = \frac{1}{N}\sum_{n=1}^{N} X_n$.

Properties of sample average II

$$
\begin{aligned}
E\overline{X} &= \frac{1}{N} E \sum_{n=1}^{N} X_n \\
\text{Why?} \quad &= \frac{1}{N} \sum_{n=1}^{N} E X_n \\
\text{Why?} \quad &= \mu_B .
\end{aligned}
$$

Properties of sample average III

The variance of the sample average

$$Var(\overline{X}) = Var(\frac{1}{N}\sum_{n=1}^{N} X_n) = \frac{1}{N^2} Var(\sum_{n=1}^{N} X_n)$$

$$\text{Independence: } = \frac{1}{N^2}\sum_{n=1}^{N} Var(X_n) = \frac{\sigma_B^2}{N}$$

And $SD(\overline{X}) = \sqrt{Var(\overline{X})} = \sigma/\sqrt{N}$.

So the expected value of the average of the sample is the average of the box no matter how large the sample.

The variance of the sample average *decreases* as the sample size N increases $\rightarrow$ The Law of Large Numbers.

Properties of sample average IV

Example: $X_1, \ldots, X_N$ are draws from a box of 0's and 1's with replacement. Assume fraction of 1's in the box is p.

Each X_n is Ber(p), i.e. $EX_n = p, Var(X_n) = p(1-p)$.

Since the draws are with replacement we can assume they are independent and so writing $S = \sum_{n=1}^{N} X_n$ we have that S is Binomial(N,p).

$f_S(n) = \binom{N}{n} p^n (1-p)^{N-n}$.

We can compute ES using the definition $\sum_{n=0}^{N} n \binom{N}{n} p^n (1-p)^{N-n}$ but that requires some complicated algebra.

Properties of sample average V

Instead we use the rules for mean and variance of a sum:
$E(S) = \sum_{n=1}^{N} EX_n = Np.$

And since X_n are independent:
$Var(S) = \sum_{n=1}^{N} Var(X_n) = Np(1-p)$

And for the sample average:
$E\overline{X} = E\frac{S}{N} = \frac{1}{N}ES = p.$

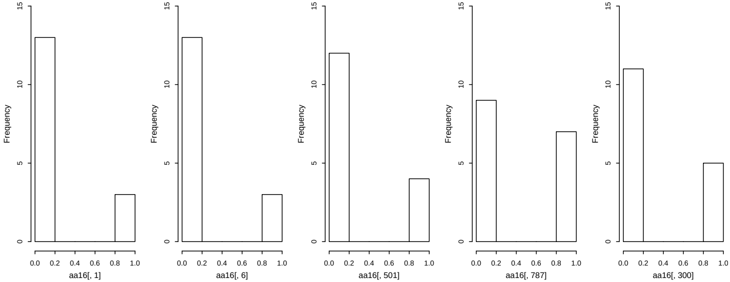
$Var(\overline{X}) = Var(\frac{S}{N}) = \frac{1}{N^2}Np(1-p) = \frac{p(1-p)}{N}.$
$SD(\overline{X}) = \frac{\sqrt{p(1-p)}}{\sqrt{N}}.$

The expected value of the sample average is p - the proportion of 1's in the box: it is centered in the 'right' place.
The variance of the sample average decreases with sample size.

Central limit theorem visual I

0/1 box with six 0's and two 1's. Get 5000 samples of size 16 with replacement. For each draw X we have $P(X = 1) = 1/4$. We show 5 of the samples. They are not identical - the sample is random.

```
ll=c(0,0,0,0,0,0,1,1)
P=sum(ll)/8
aa16=replicate(5000,sample(ll,16,replace=TRUE))
```



Central limit theorem visual II

Now draw 5000 samples of size 100, and of size 10000. Compute the sums of the samples

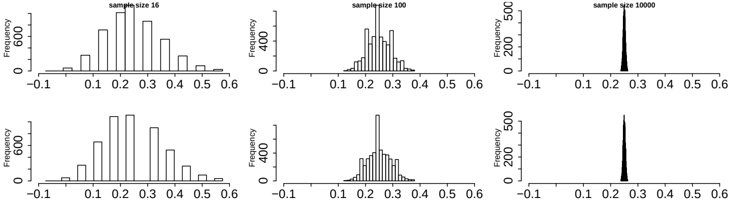
```
aa100=replicate(5000,sample(ll,100,replace=TRUE))
aa10000=replicate(5000,sample(ll,10000,replace=TRUE))
sumaa16=colSums(aa16)
sumaa100=colSums(aa100)
sumaa10000=colSums(aa10000)
```

Same as sampling 5000 times from binomial distributions $Bin(16, 1/4), Bin(100, 1/4), Bin(10000, 1/4)$ respectively:

```
b16=rbinom(5000,16,P)
b100=rbinom(5000,100,P)
b10000=rbinom(5000,10000,P)
```

Central limit theorem visual III

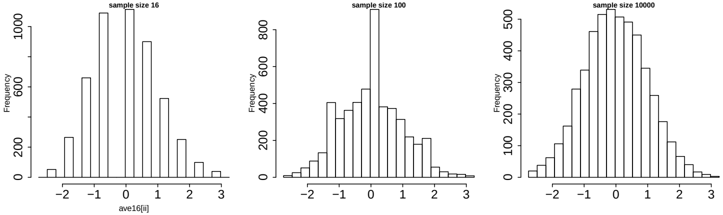
Show the histograms of sample averages: divide the sums by (16,100,10000) - top row, and divide the binomials by (16,100,10000) - bottom row. They look the same *because they are draws from the same distribution.*



Notice how the spread of the histograms gets smaller and smaller but they are all centered very close to $1/4$.
Law of Large Numbers.

Central limit theorem visual IV

Now, instead standardize the lists of averages: substract the expected value of the average $p = 1/4$ and divide by SDs of the average $\sqrt{(3/16)/16}$, $\sqrt{(3/16)/100}$, $\sqrt{(3/16)/10000}$



Now the histograms all have the same spread and are centered at zero, but they are looking more and more like the normal distribution.

The normal density I

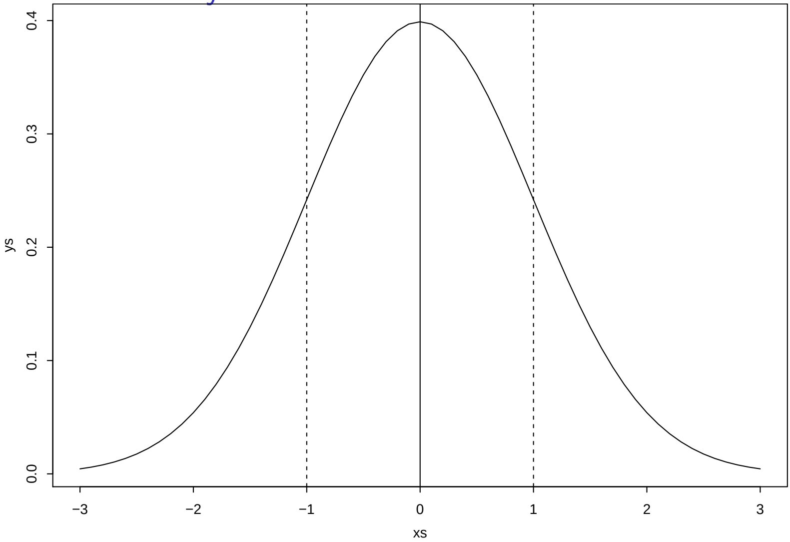
Standard Normal density:

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

A random variable X with standard normal density has:

$$EX = \int xf(x)dx = 0, \quad Var(X) = \int (x - \mu_X)^2 f(x) = 1.$$

The normal density II



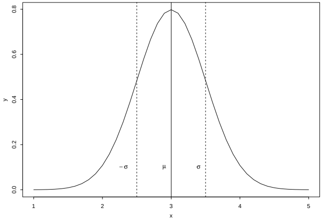
The normal density III

Normal density with mean μ and variance σ^2

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

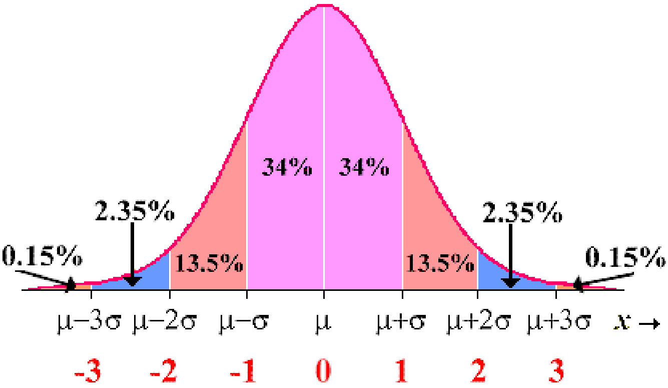
A random variable X with normal density $f(x; \mu, \sigma)$ has:

$$\mu_X = EX = \mu, \quad \sigma_X^2 = Var(X) = \sigma^2.$$



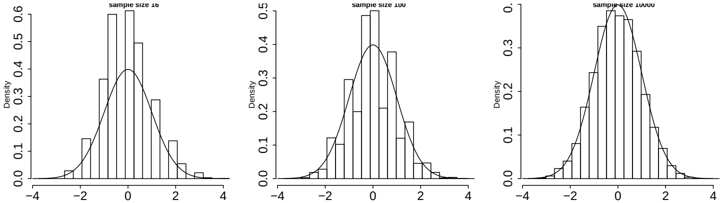
The normal density IV

The 68-95-99.7 rule for the normal distribution.



The normal density V

Compare historgram of **standardized** sample averages to normal density



Check if sample is normal I

How do data compare to normal distribution.

First standardize data: $z_i = (x_i - \bar{x})/sd(x)$.

Then compare quantiles: what percentage of z_i's are below -1 vs percentage of standard normal below -1 which is about .16.

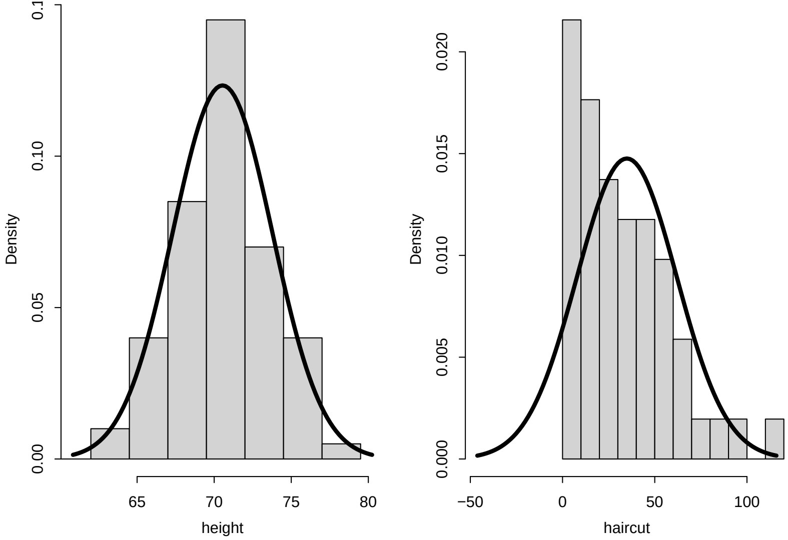
What percentage of standardized data are below 1 vs percentage of normal below -1 which is about .84.

We can do that for multiple quantiles and create pairs. If Φ is CDF of normal:

$x(q) = \Phi^{-1}(q), z(q) =$ q'th quantile of data.

Check if sample is normal II

Data on height of males in a class and amount spent on haircuts.



Check if sample is normal III

First standardize:

```
stdHaircut = (haircut - mhair) / shair
stdheight = (height-mht)/sht
```

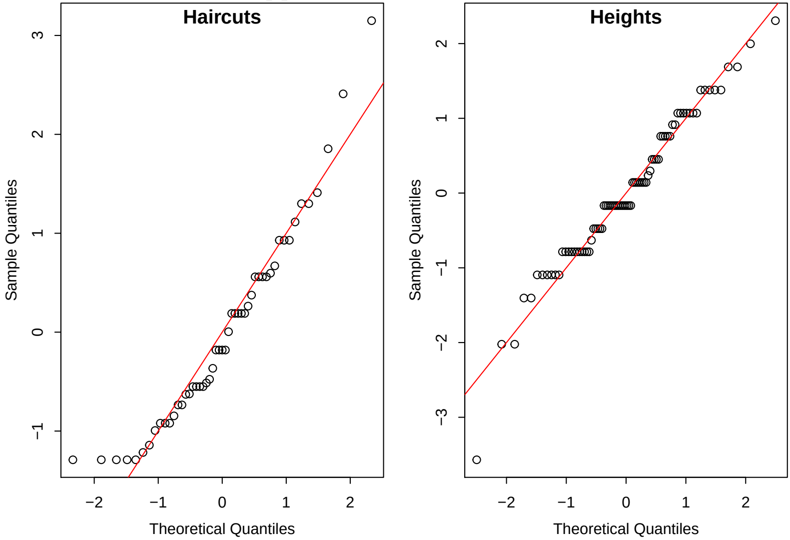
Then compare quantiles:

```
r = c(0.0015, 0.025, 0.16, 0.25, 0.50, 0.75, 0.84, 0.975, 0.9985
modelQuantile = qnorm(r)
dataQuantile = quantile(stdHaircut, r, na.rm=TRUE)
rbind(dataQuantile, modelQuantile)
```

```
               0.15%  2.5%     16%     25%     50%    75%
dataQuantile  -1.290 -1.29 -0.9205 -0.7355 -0.1805 0.5594
modelQuantile -2.968 -1.96 -0.9945 -0.6745  0.0000 0.6745
                 84% 97.5% 99.85%
dataQuantile  0.9294 2.271  3.094
modelQuantile 0.9945 1.960  2.968
```

Check if sample is normal IV

Or let R do it for you: `qqnorm(stdHaircut)`



Check if sample is normal V