Lecture 8 - Normal Approximation to the Binomial

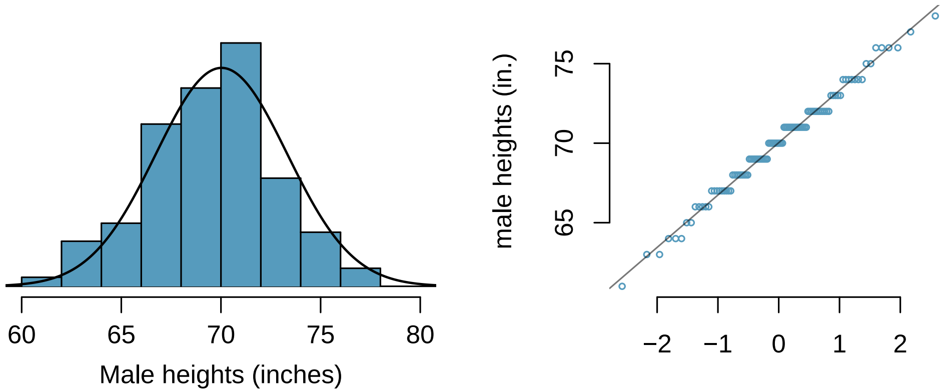
Sta102 / BME102

Colin Rundel

September 21, 2015

Normal probability plot

Below is a histogram, with a superimposed normal distribution, of a sample of 100 male heights. Does height appear to be normally distributed?



Anatomy of a normal probability plot

- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis.
- If there is a linear relationship between the data and the theoretical quantiles, then the data follow a nearly normal distribution.
- Since a linear relationship should appear as a straight line on the scatter plot, the closer the points are to a perfect straight line, the more confident we can be that the data follow the normal distribution.
- Constructing a normal probability plot requires calculating percentiles and corresponding Z-scores for each observation, which is tedious. Therefore we generally rely on software when making these plots.

Constructing a normal probability plot

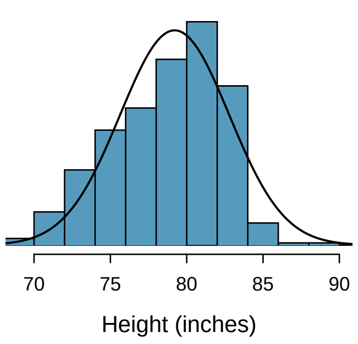
We construct a normal probability plot for the heights of a sample of 100 men as follows:

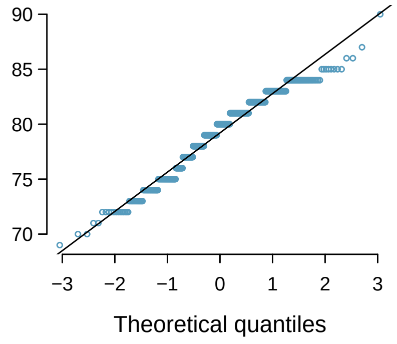
1. Order the observations.
2. Determine the percentile of each observation in the ordered data set.
3. Identify the Z score corresponding to each percentile (using a Z table).
4. Create a scatterplot of the observations (y) against the Z scores (x)

Observation i	1	2	3	$\cdots$	100
x_i	61	63	63	$\cdots$	78
Percentile	1%	2%	3%	$\cdots$	99%
Z_i	-2.33	-2.06	-1.89	$\cdots$	2.33

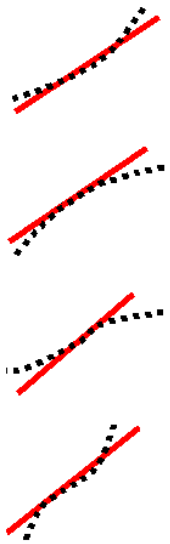
Example - NBA Height

Below is a histogram and normal probability plot for the heights of NBA players. Do these data appear to follow a normal distribution?





Normal probability plot and skewness

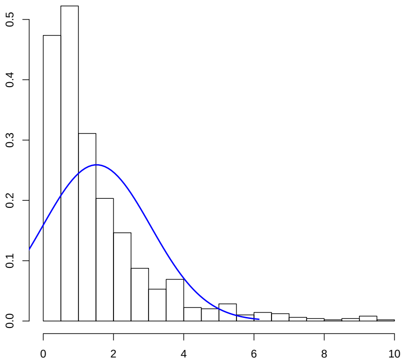


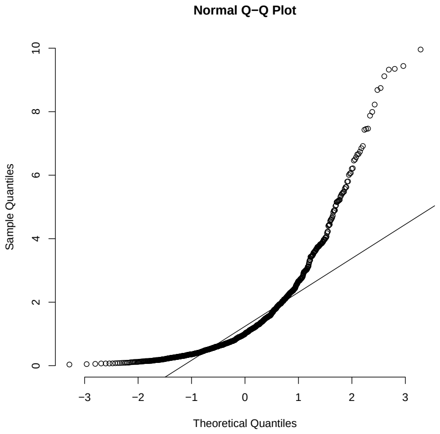
Right Skew - If the plotted points appear to bend up and to the left of the normal line that indicates a long tail to the right.

Left Skew - If the plotted points bend down and to the right of the normal line that indicates a long tail to the left.

Long/Fat Tails - A curve which starts below the normal line, bends to follow it, and ends above it indicates long tails. That is, you are seeing more variance than you would expect in a normal distribution, i.e. wider than expected.
Short/Skinny Tails - An S shaped-curve indicates shorter than normal tails, i.e. narrower than expected.

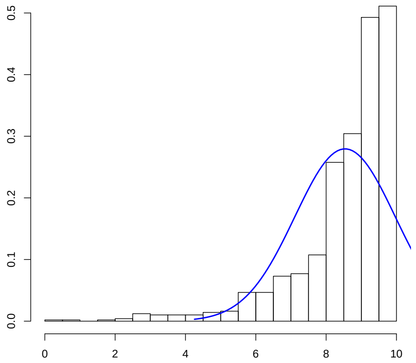
Right Skew

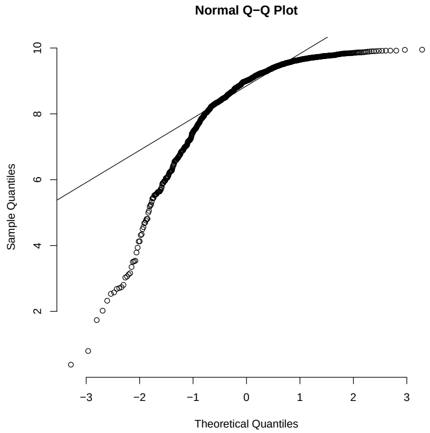




Here the biggest values are bigger than we would expect and the smallest values are also bigger than we would expect.

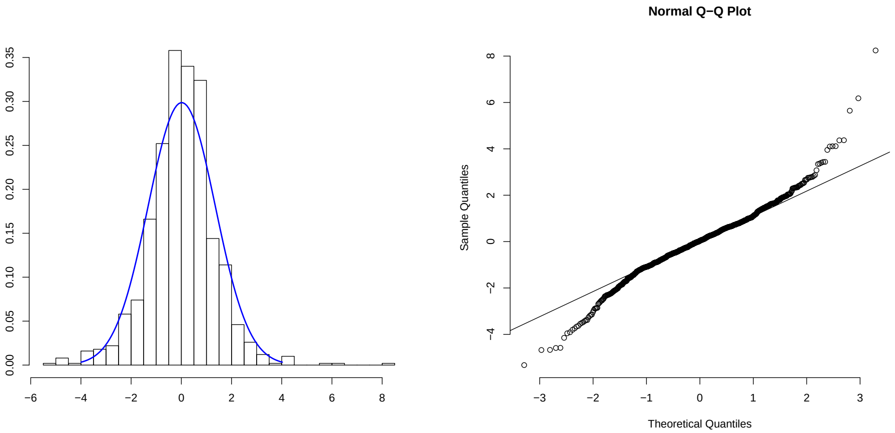
Left Skew





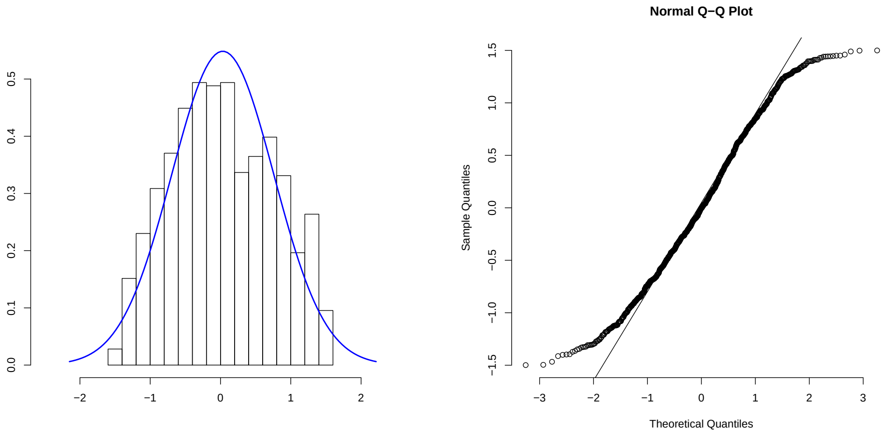
Here the biggest values are smaller than we would expect and the smallest values are also smaller than we would expect.

Fat tails



Best to think about what is happening with the most extreme values - here the biggest values are bigger than we would expect and the smallest values are smaller than we would expect (for a normal).

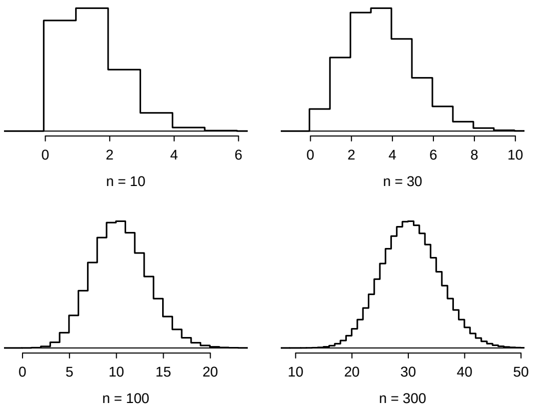
Skinny tails



Here the biggest values are smaller than we would expect and the smallest values are bigger than we would expect.

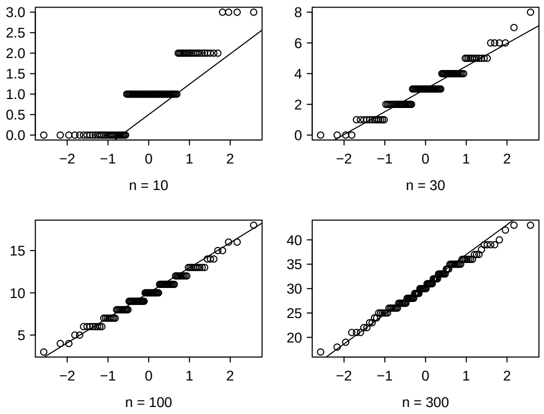
Histograms of the number of successes

Hollow histograms of samples from a binomial model where $p = 0.10$ and $n = 10$, 30, 100, and 300. What happens as n increases?



QQ plots of the number of successes

QQ plots of samples from a binomial model where $p = 0.10$ and $n = 10$, 30, 100, and 300. What happens as n increases?



In general, if $np \geq 10$ and $n(1-p) \geq 10$ then approximately normal.

de Moivre-Laplace Limit Theorem

When n is large enough the Binomial distribution will always have this bell-curve shape.

- Approximation is usually considered reasonable when $np \geq 10$ and $n(1-p) \geq 10$

de Moivre and Laplace where the first to identify this pattern (in the 18th century) and characterize the shape of the curve.

This is a special case of a more general result known as the Central Limit Theorem. (More on this on Wednesday)

Example - Drosophila

A geneticist is studying a population of *Drosophila* where 25% of the flies have white eyes, the other 75% have red eyes. For an upcoming experiment the scientist needs at least 70 white eyed flies. If they are able to collect 245 larvae what is the probability that they will have sufficient white eyed flies for their experiment?

We are given that $n = 245, p = 0.25$, and we are asked to find the probability $P(X \geq 70)$.

$$\begin{aligned} P(X \geq 70) &= P(X = 70 \text{ or } X = 71 \text{ or } X = 72 \text{ or } \cdots \text{ or } X = 245) \\ &= P(X = 70) + P(X = 71) + P(X = 72) + \cdots + P(X = 245) \end{aligned}$$

This seems like an awful lot of work...

Normal approximation to the binomial

When the number of trials (n) is large enough, a binomial distribution (X) has an approximately normal distribution (X') where

$$\mu = E(X) = np \quad \text{and} \quad \sigma = SD(X) = \sqrt{np(1-p)}.$$

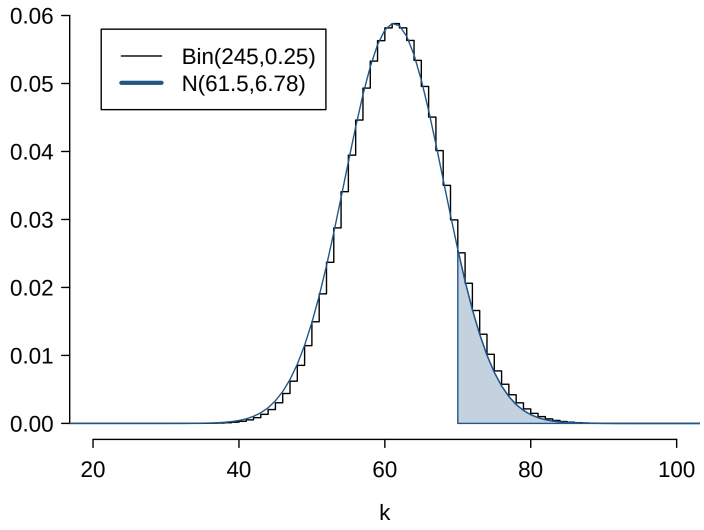
- For our *Drosophila* experiment, $n = 245$ and $p = 0.25$.

  $E(X) = 245 \times 0.25 = 61.25 \qquad SD(X) = \sqrt{245 \times 0.25 \times 0.75} = 6.78$

- As such, for the probability $P(X \geq x)$ we can approximate it using $P(X' \geq x)$ where

  $X \sim \text{Binom}(n = 245, p = 0.25) \quad \text{and} \quad X' \sim N(\mu = 61.25, \sigma = 6.78).$

Normal approximation to the binomial (graphically)



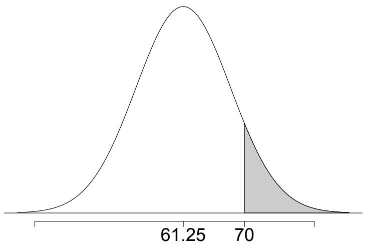
Drosophila cont.

What is the probability that among the 245 larvae there are 70 or more white eyed genotypes?

Let $X \sim \text{Binom}(n = 245,\ p = 0.25)$ and $X' \sim N(\mu = 61.25,\ \sigma = 6.78)$ then

$$P(X \geq 70) \approx P(X' \geq 70) = ?P(Z \geq 1.29) = 0.0985$$

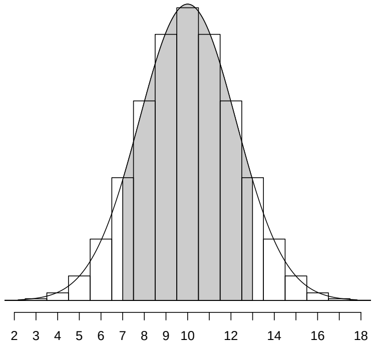
$$Z = \frac{x - E(X)}{SD(X)} = \frac{70 - 61.25}{6.78} = 1.29$$



Z	0.05	0.06	0.07	0.08	0.09
1.0	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8944	0.8962	0.8980	0.8997	0.9015

Improving the approximation

Take for example the Binomial distribution $X \sim \text{Binom}(n = 20, p = 0.5)$, we should be able to approximate this distribution using $X' \sim N(10, \sqrt{5})$.



Our approximation is missing about 1/2 of $P(X = 7)$ and $P(X = 13)$, which is $\approx 7\%$. (This error shrinks as n increases)

Improving the approximation, cont.

Binomial probability:

$$P(7 \leq X \leq 13) = \sum_{k=7}^{13} \binom{20}{k} 0.5^k (1 - 0.5)^{20-k} = 0.88468$$

Naive approximation:

$$P(7 \leq X \leq 13) \approx P\left(Z \leq \frac{13 - 10}{\sqrt{5}}\right) - P\left(Z \leq \frac{7 - 10}{\sqrt{5}}\right) = 0.82029$$

Continuity corrected approximation:

$$P(7 \leq X \leq 13) \approx P\left(Z \leq \frac{13 + 1/2 - 10}{\sqrt{5}}\right) - P\left(Z \leq \frac{7 - 1/2 - 10}{\sqrt{5}}\right) = 0.88248$$

Improving the approximation, cont.

This correction also lets us do clever things like calculate the probability for a particular value of k. Such as, what is the chance of 50 Heads in 100 tosses of slightly unfair coin ($p = 0.55$)?

Binomial probability:

$$P(X = 50) = \binom{100}{50} 0.55^{50} (1 - 0.55)^{50} = 0.04815$$

Naive approximation:

$$P(X = 50) \approx P\left(Z \leq \frac{50 - 55}{4.97}\right) - P\left(Z \leq \frac{50 - 55}{4.97}\right) = 0$$

Continuity corrected approximation:

$$P(X = 50) \approx P\left(Z \leq \frac{50 + 1/2 - 55}{\sqrt{4.97}}\right) - P\left(Z \leq \frac{50 - 1/2 - 55}{\sqrt{4.97}}\right) = 0.04839$$

Example - Rolling lots of dice

Roll a fair die 500 times, what's the probability of rolling at least 100 ones?

$$\begin{aligned}
P(X \geq 100) &= \sum_{k=100}^{500} \binom{500}{k} (1/6)^k \; (5/6)^{500-k} \\
&= 1 - \sum_{k=0}^{99} \binom{500}{k} (1/6)^k \; (5/6)^{500-k} \\
&= 1 - \texttt{pbinom}(99, 500, 1/6) \\
&= 1 - 0.9717129 \\
&= 0.0282871
\end{aligned}$$

Example - Rolling lots of dice

Roll a fair die 500 times, what's the probability of rolling at least 100 ones?

Since n is large, X can be approximated with a normal distribution (X') where $\mu = E(X) = np = 500/6 = 83.33$ and $\sigma = SD(X) = \sqrt{npq} = \sqrt{2500/36} = 8.333$

$$\begin{aligned}
P(X \geq 100) \approx & P(X' \geq 100) \\
&= P\left(Z \geq \frac{100 - 1/2 - \mu}{\sigma}\right) \\
&= P\left(Z \geq \frac{100 - 1/2 - 83.33}{8.333}\right) \\
&= 1 - P(Z \leq 1.94) \\
&= 1 - 0.9738 \\
&= 0.0262
\end{aligned}$$

Example - Airline booking

An airline knows that over the long run, 90% of passengers who reserve seats show up for flight. On a particular flight with 300 seats, the airline accepts 324 reservations. If passengers show up independently what is the probability the flight will be overbooked?

Example - Voter support

Suppose 55% of a large population of voters support actually favor a particular candidate. How large a random sample must be take for there to be a 99% chance that the majority of voters in the sample will favor that candidate?

Example - Roulette

Suppose you enter a casino and plan to play roulette by betting $1 on black for every spin. Assuming you do this for 8 hours and the croupier spins the wheel once a minute. What is the probability that you break even or come out ahead? (Win as many times or more than you lose.)