Lecture 6 - Assessing Normality, Normal Approximation to Binomial

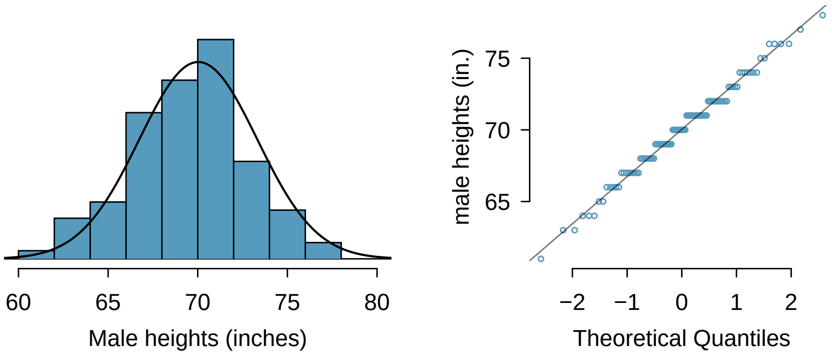
Sta102/BME102

Colin Rundel

February 5, 2013

Normal probability plot

A histogram and *normal probability plot* of a sample of 100 male heights.



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 one-to-one relationship between the data and the theoretical quantiles, then the data follow a nearly normal distribution.
- Since a one-to-one relationship would appear as a straight line on a scatter plot, the closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model.
- 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 as follows:

1. Order the observations.
2. Determine the percentile of each observation in the ordered data set $\left(\text{Percentile of the } i^{th} \text{ observation is } P_i = \frac{i}{n+1}\right)$.
3. Identify the Z score corresponding to each percentile.
4. Create a scatterplot of the observations (vertical) against the Z scores (horizontal)

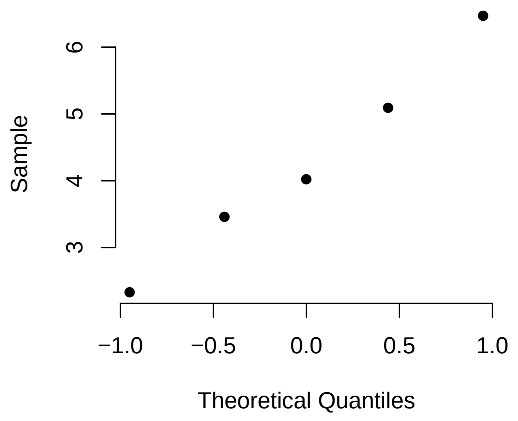
QQ Example

Does the following data have an approximately normal distribution?

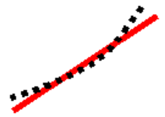
3.46, 4.02, 5.09, 2.33, 6.47

Obs	1	2	3	4	5
y_i	2.33	3.46	4.02	5.09	6.47
P_i	0.17	0.33	0.50	0.67	0.83
Z_i	-0.95	-0.44	0	0.44	0.95

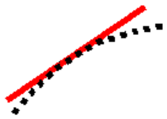
QQ Example



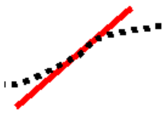
Normal probability plot and skewness



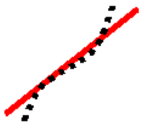
Right Skew - If the plotted points appear to bend up and to the left of the normal line that indicates a long tail to the right.



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

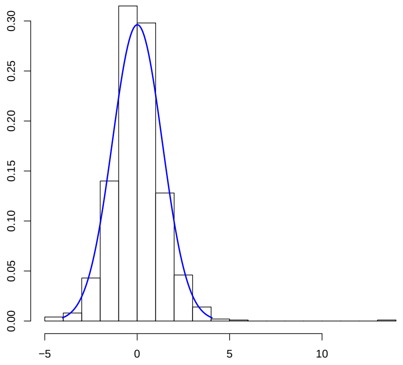


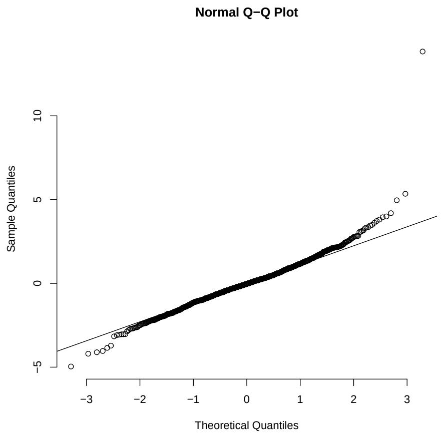
Short/Skinny Tails - An S shaped-curve indicates shorter than normal tails, i.e. narrower than expected.



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.

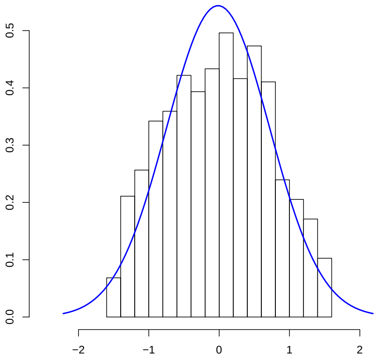
Fat tails

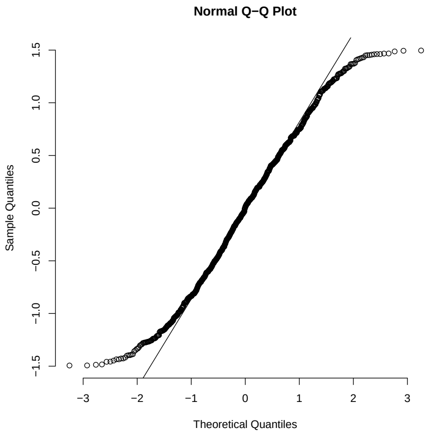




Best to think about what is happening with the most extreme values - here the biggest values are bigger than we would expect and the smallest values are smaller than we would expect (for a normal).

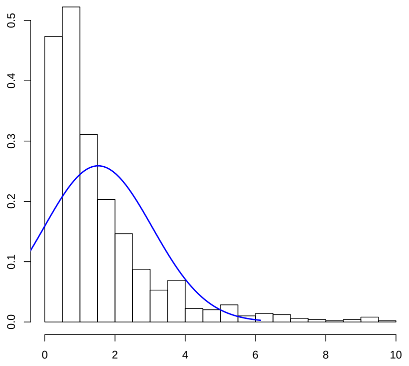
Skinny tails

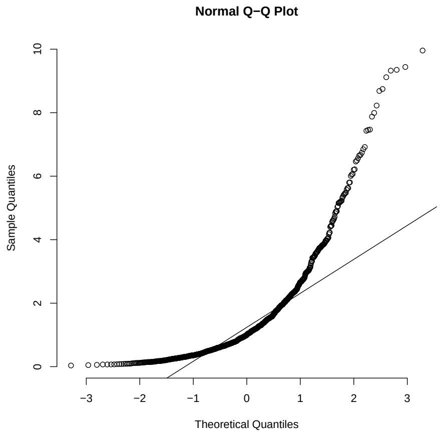




Here the biggest values are smaller than we would expect and the smallest values are bigger than we would expect.

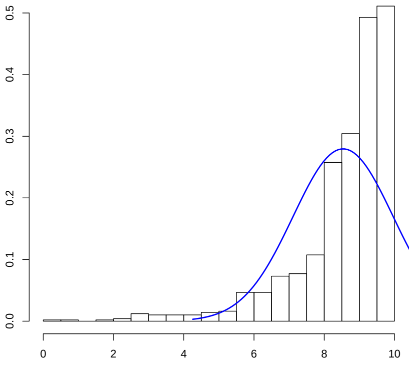
Right Skew

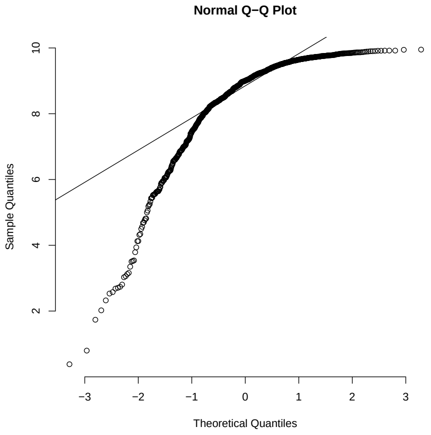




Here the biggest values are bigger than we would expect and the smallest values are also bigger than we would expect.

Left Skew

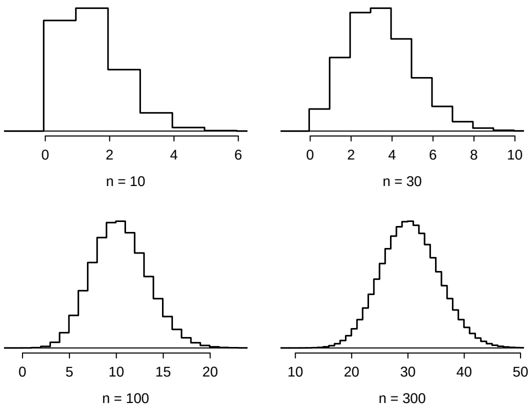




Here the biggest values are smaller than we would expect and the smallest values are also smaller than we would expect.

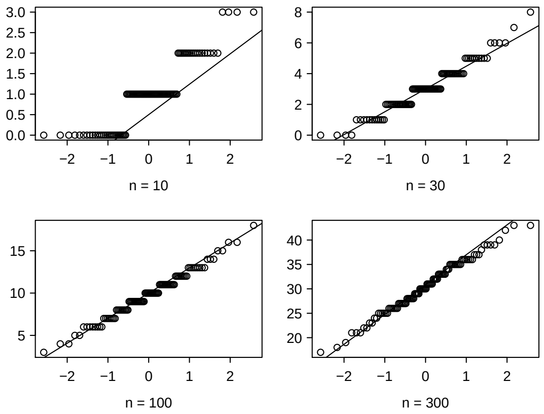
Histograms of the number of successes

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 normal approximation is reasonable.

An analysis of Facebook users

A recent study found that "Facebook users get more than they give". For example:

- 40% of Facebook users in our sample made a friend request, but 63% received at least one request
- Users in our sample pressed the like button next to friends' content an average of 14 times, but had their content "liked" an average of 20 times
- Users sent 9 personal messages, but received 12
- 12% of users tagged a friend in a photo, but 35% were themselves tagged in a photo

Any guesses for how this pattern can be explained?

http://www.pewinternet.org/Reports/2012/Facebook-users/Summary.aspx

Facebook cont.

This study found that approximately 25% of Facebook users are considered power users. The same study found that the average Facebook user has 245 friends. What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?

We are given that $n = 245, p = 0.25$, and we are asked for 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...

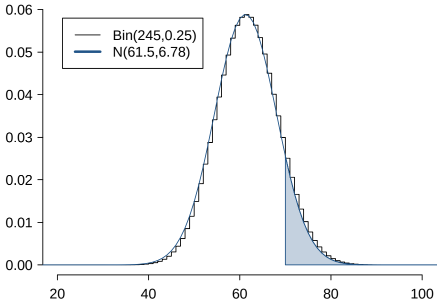
Normal approximation to the binomial

When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal model with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

- In the case of the Facebook power users, $n = 245$ and $p = 0.25$.

$$\mu = 245 \times 0.25 = 61.25 \qquad \sigma = \sqrt{245 \times 0.25 \times 0.75} = 6.78$$

- $\text{Binom}(n = 245, p = 0.25) \approx N(\mu = 61.25, \sigma = 6.78)$.

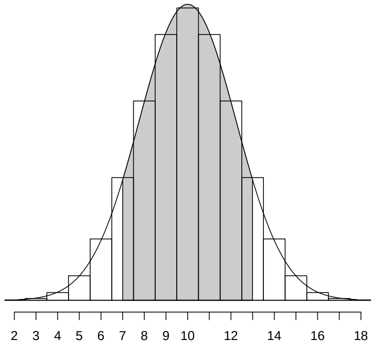


Facebook cont.

What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?

Improving the approximation

Take for example a Binomial distribution where $n = 20$ and $p = 0.5$, we should be able to approximate the distribution of X using $N(10, \sqrt{5})$.



It is clear that our approximation is missing about $1/2$ of $P(X = 7)$ and $P(X = 13)$, as $n \to \infty$ this error is very small. In this case $P(X = 7) = P(X = 13) = 0.073$ so our approximation is off by $\approx 7\%$.

Improving the approximation, cont.

Binomial probability:

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

Naive approximation:

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

Continuity corrected approximation:

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

Improving the approximation, cont.

This correction also lets us do, moderately useless, 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:

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

Continuity corrected approximation:

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

Example - Rolling lots of dice

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

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?

Suppose that people travel in groups, does this increase or decrease the chance of overbooking?

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?