Stat 13, Thu 5/3/12.

1. Correction on normal notation.
2. Normal percentiles.
3. Normal probability plots.
4. Bernoulli and Binomial random variables.
5. E(x+y).
6. Independence and alleles.

Hw4 is due Tue, 5/8. Midterm 2 is Thur, 5/17.
When submitting hw, please make 4 stacks: section 1a, 1b, 1c, & 1d.

Ignore the chapter on the "continuity correction".

Gradegrubbing procedure: if you would like a question (or more than one question) reevaluated, submit your exam or homework and a **WRITTEN** explanation of why you think you deserve more points and how many more points you think you deserve ***to your TA.*** The TA will then give it to me, and I will consider it, and then give it back to the TA to give back to you.

1. Correction on normal notation.
Last time I said N(μ, σ^2) means a normal random variable with mean μ and sd σ.
However, in ch4, your book just calls this N(μ, σ). I will use the book's notation from now on, and I actually changed day7.ppt since last class, so it is now consistent with this as well.

2. Normal percentiles.
The main point from last time was that if X is N(μ, σ), and Z = $\frac{X-\mu}{\sigma}$, then Z is N(0,1), i.e. standard normal.
So, P(X < c) = P($\frac{X-\mu}{\sigma} < \frac{c-\mu}{\sigma}$) = P(Z < $\frac{c-\mu}{\sigma}$), which you can find in the table in your book.

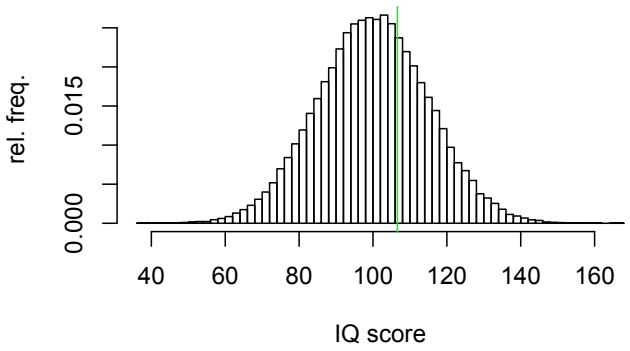
Now, how about if, instead of P(X < c) , you want to find c such that P(X < c) = k%?
That is, suppose you want to find the kth percentile of your distribution?
For instance, IQs are N(100, 15). What is your IQ if you are in the 90th percentile?
We want c, where 90% of people score less than c.
That is, if X is a randomly chosen person,
then we want c so that P(X < c) = 90%.
Look in the body of the table til you see 90%. Find P(Z < 1.28) = 0.8997 ~ 90%.
So, 90% = P(Z < 1.28) = P($\frac{X-\mu}{\sigma} < 1.28$) = P(X - μ < 1.28 σ) = P(X < 1.28σ + μ)
= P(X < 1.28(15) + 100)
= P(X < 119.2).
Thus, the answer is 119.2. Previously, to convert *to* standard units, c -> (c-μ)/σ, and here we know what the answer c is in standard units, so to convert c *from* standard units to IQ points, we do the opposite, i.e. **c -> cσ + μ**.

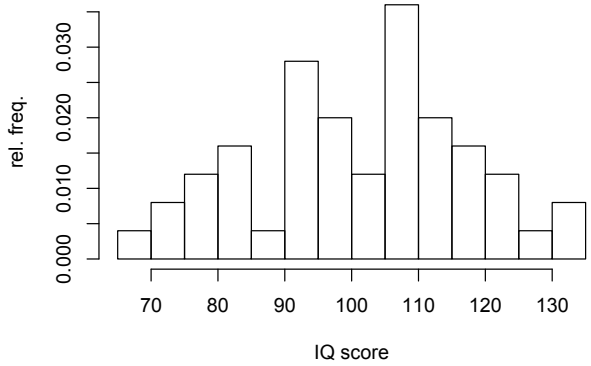
3. Normal probability plots.

How can you tell if a distribution is normal, or approximately normal?

You could look at the histogram and see if it looks like the normal curve.

With enough data, this isn't a problem. However, with less data, it can be hard to tell.



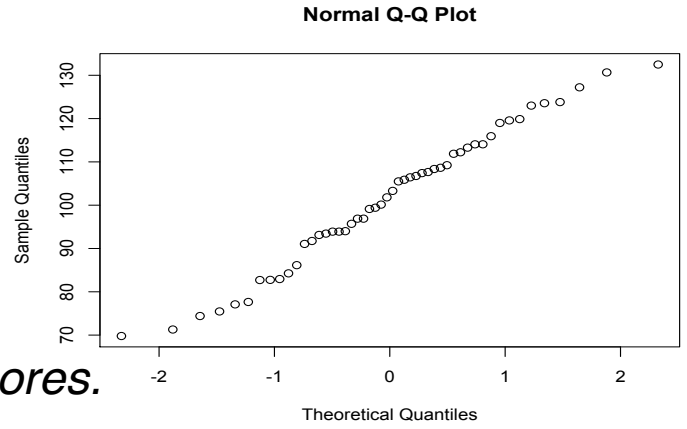


An alternative approach is the normal probability plot. It's also called a normal Q-Q plot. The Qs stand for quantile.

Sort the data, and for data point X_i, it's the i/n quantile of your data.

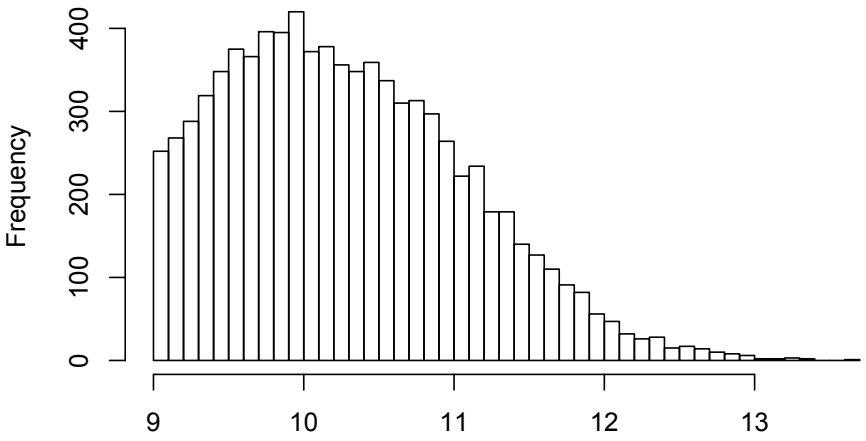
Figure out what you'd expect the i/n quantile of the standard normal to be, call it q_i, and plot X_i vs. q_i.

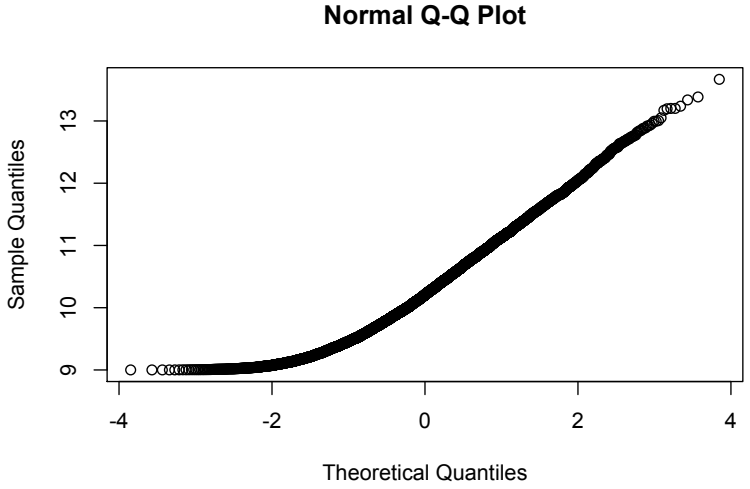
(In most computer packages, the default is actually to plot q_i vs. X_i. Book calls q_i *n-scores.*

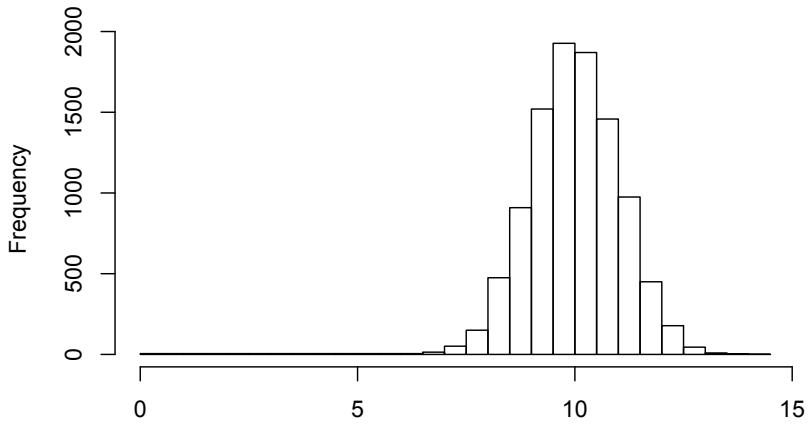


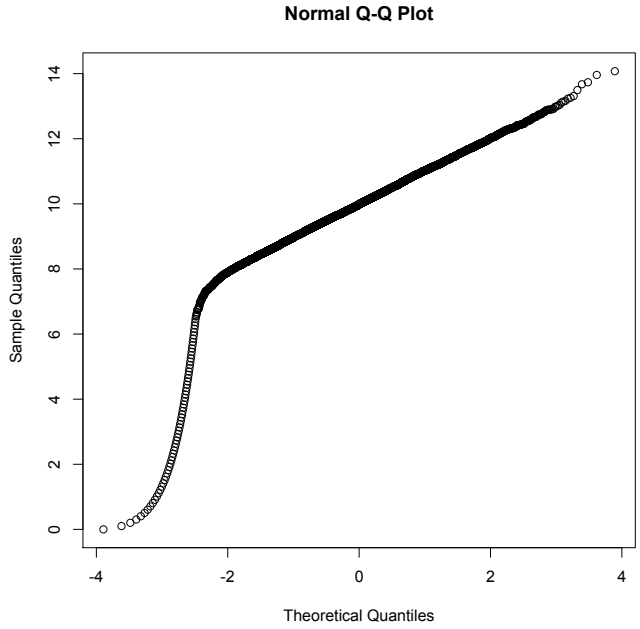
Linear means good fit.

Curvature typically means one or both tails of the distribution are non-normal.


Frequency
0
100
200
300
400
9
10
11
12
13


Normal Q-Q Plot
Sample Quantiles
9
10
11
12
13
-4
-2
0
2
4
Theoretical Quantiles


Frequency
0
500
1000
1500
2000
0
5
10
15


Normal Q-Q Plot
Sample Quantiles
0
2
4
6
8
10
12
14
-4
-2
0
2
4
Theoretical Quantiles

4. Bernoulli and Binomial random variables.

Often a variable has only 2 possible outcomes, like a coin flip, or people's genders, or if your response variable is whether someone has a disease or not.
A useful convention is to score the responses 0 or 1. (not -1 or 1). Such random variables are then called Bernoulli, or Bernoulli (p), if p = probability of scoring 1.

If X is Bernoulli (p), then E(X) = p, and SD(X) = $\sqrt{pq}$.

Now suppose you flip n coins, or sample n subjects and record their genders.
Sometimes you are interested in the *total* number of 1s in your sample.
The 0-1 convention is helpful, because this total number is the same as the sum of your sample values.
If each obs. is 0 or 1, and Y = the total number of 1s in your sample, and if the observations are independent, then we say Y is a binomial(n,p) random variable.
P(Y=k) is C(n,k) p^k q^{n-k}, where q = 1-p, for k = 0, 1, 2, ..., n.
For instance, flip 8 coins. P(Y=3) = P(HHHTTTTT or HTHTHTTT or ...). There are C (8,3) different places you could put the H's, and each such ordering has prob. p^3 q^5.
If Y is binomial(n,p), then E(Y) = np, and SD(Y) = $\sqrt{npq}$.

5. E(X+Y).

In general, E(X+Y) = E(X) + E(Y), even if X and Y might *not* be independent!

For example, suppose you bet $10 on number 32 in roullette.
Your expected profit is 10 x -5.3 cents = -53 cents.
Now suppose instead you put $5 on number 32 and $5 on 33.
Your total expected profit from the two bets is
E(profit on 32 + profit on 33) = E(profit on 32) + E(profit on 33)
= 5 x -5.3 + 5 x -5.3 = -53 cents.
Now suppose you put $5 on number 32 on one spin, and then $5 on 33 on the next spin. Your total expected profit is still
E(profit on 32 + profit on 33) = E(profit on 32) + E(profit on 33)
= 5 x -5.3 + 5 x -5.3 = -53 cents.

6. Independence and alleles.

People often multiply allele frequencies to estimate the probability of a match. This assumes independence of alleles!
P(randomly selected person matches allele #1 and allele #2 and ...
and allele #6)
= P(randomly selected person matches allele #1) x P(randomly selected person matches allele #2) x ... x P(randomly sel. person matches #6)
= 12/240 x 37/750 x ... This assumes independence!
For instance, in the OJ Simpson trial, Bruce Weir multiplied probabilities to claim that there was a 1 in 140 million chance that a randomly selected person's blood would have matched the blood found at the crime scene in all the tested alleles. The allele frequencies, however, were based on samples of just a few hundred people.
Don't assume independence unless there is a good reason to do so!
(coins, dice, roullette spins, sampling with replacement, sampling from a large population, can verify that the events are independent because P(AB) = P(A)P(B), or you are told that the events or variables are independent).