Random variables

9.07

2/19/2004

A few notes on the homework

- If you work together, tell us who you're working with.
  - You should still be generating your own homework solutions. Don't just copy from your partner. We want to see your own words.
- Turn in your MATLAB code (this helps us give you partial credit)
- Label your graphs
  - xlabel('text')
  - ylabel('text')
  - title('text')

More homework notes

- Population vs. sample
  – The population to which the researcher wants to generalize can be considerably more broad than might be implied by the narrow sample.
    - High school students who take the SAT
    - High school students
    - Anyone who wants to succeed
    - Anyone

More homework notes

- MATLAB:
  - If nothing else, if you can't figure out something in MATLAB, find/email a TA, or track down one of the zillions of fine web tutorials.
  - Some specifics…

MATLAB

- Hint: MATLAB works best if you can think of your problem as an operation on a matrix. Do this instead of “for” loops, when possible.
  - E.G. coinflip example w/o for loops

```
x = rand(5,10000);
coinflip = x>0.5;
numheads = sum(coinflip);   % num H in 5 flips
```

MATLAB

- randn(N) -> NxN matrix!
- randn(1,N) -> 1xN matrix
- sum(x) vs. sum(x,2)
- hist(data, 1:10) vs. hist(data, 10)
- plot(hist(data)) vs.
  [n,x]=hist(data); plot(x,n)

A few more comments

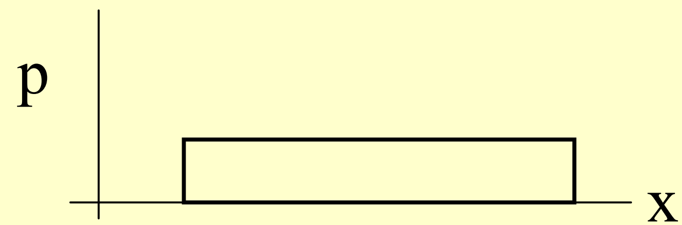
- Expected value can tell you whether or not you want to play game even once.
  - It tells you if the "game" is in your favor.
- In our example of testing positive for a disease, P(D) is the *prior* probability that you have the disease. What was the probability of you having the disease before you got tested? If you are from a risky population, P(D) may be higher than 0.001. Before you took the test you had a higher probability of having the disease, so after you test positive, your probability of having the disease, P(D|+) will be higher than 1/20.

Random Variables

- Variables that take numerical values associated with events in an experiment
  - Either discrete or continuous
    - Integral (not sum) in equations below for continuous r.v.
  - Mean, μ, of a random variable is the sum of each possible value multiplied by its probability: $\mu = \sum x_i P(x_i) \equiv E(x)$
    - Note relation to "expected value" from last time.
  - Variance is the average of squared deviations multiplied by the probability of each value
  - $\sigma^2 = \sum (x_i - \mu)^2 P(x_i) \equiv E((x-\mu)^2)$

We’ve already talked about a few special cases

- Normal r.v.’s (with normal distributions)
- Uniform r.v.’s (with distributions like this:)



- Etc.

Random variables

- Can be made out of functions of other random variables.
- X r.v., Y r.v. ->

  Z=X+Y r.v.

  Z=sqrt(X)+5Y + 2 r.v.

Linear combinations of random variables

- We talked about this in lecture 2. Here's a review, with new E() notation.
- Assume:
  - $E(x) = \mu$
  - $E(x-\mu)^2 = E(x^2-2\mu x+\mu^2) = \sigma^2$
- $E(x+5) = E(x) + E(5) = E(x) + 5 = \mu + 5 = \mu'$
- $E((x+5-\mu')^2) = E(x^2+2(5-\mu')x + (5-\mu')^2)$
  $= E(x^2-2\mu x+\mu^2) = \sigma^2 = (\sigma')^2$

Adding a constant to x adds that constant to μ, but leaves σ unchanged.

Linear combinations of random variables

- $E(2x) = 2E(x) = 2\mu = \mu'$
- $E((2x-\mu')^2) = E(4x^2 - 8x\mu + 4\mu^2) = 4\sigma^2 = (\sigma')^2$

  $\sigma' = 2\sigma$

Scaling x by a constant scales both μ and σ by that constant. But…

Multiplying by a negative constant

- $E(-2x) = 2E(x) = -2\mu = \mu'$
- $E((-2x-\mu')^2) = E(4x^2 + 2(2x)(-2\mu) + (-2\mu)^2)$
  $= E(4x^2 - 8x\mu + 4\mu^2) = 4\sigma^2 = (\sigma')^2$
  $\sigma' = 2\sigma$

Scaling by a negative number multiples the mean by that number, but multiplies the standard deviation by –(the number). (Standard deviation is always positive.)

What happens to z-scores when you apply a transformation?

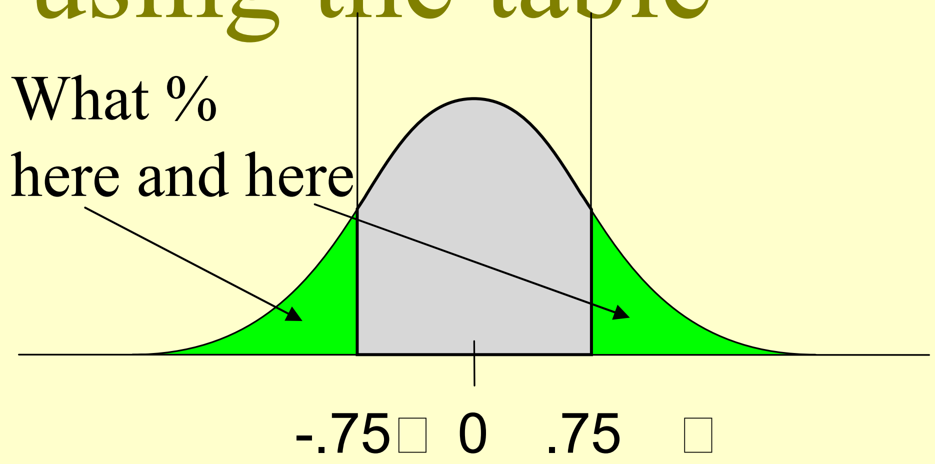
- Changes in scale or shift do not change "standard units," i.e. z-scores.
  - When you transform to z-scores, you're already subtracting off any mean, and dividing by any standard deviation. If you change the mean or standard deviation, by a shift or scaling, the new mean (std. dev.) just gets subtracted (divided out).

Special case: Normal random variables

- Can use z-tables to figure out the area under part of a normal curve.

An example of using the table

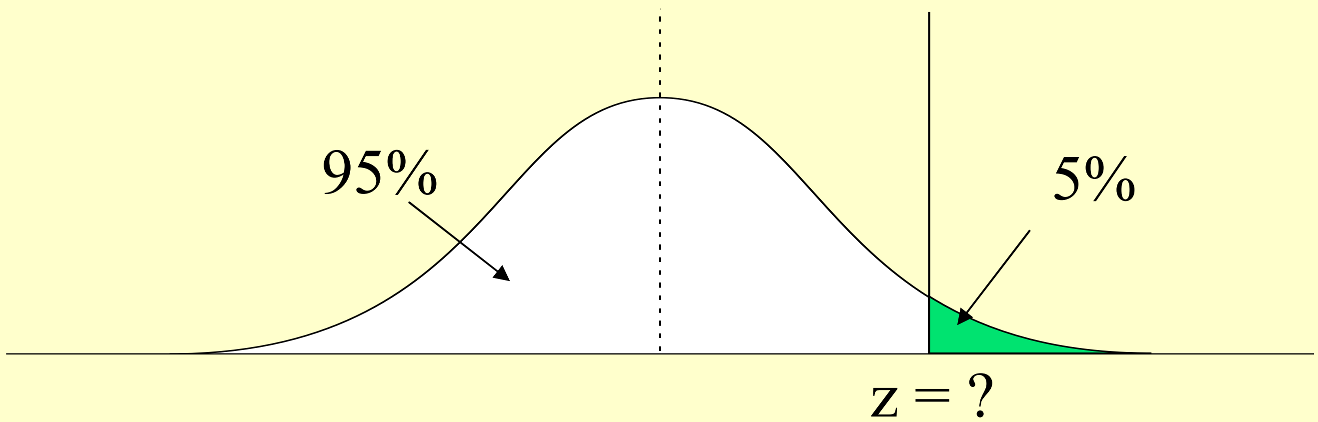
- P(-0.75<z<0.75) = 0.5467
- P(z<-0.75 or z>0.75) = 1-0.5467 ≈ 0.45
- That's our answer.



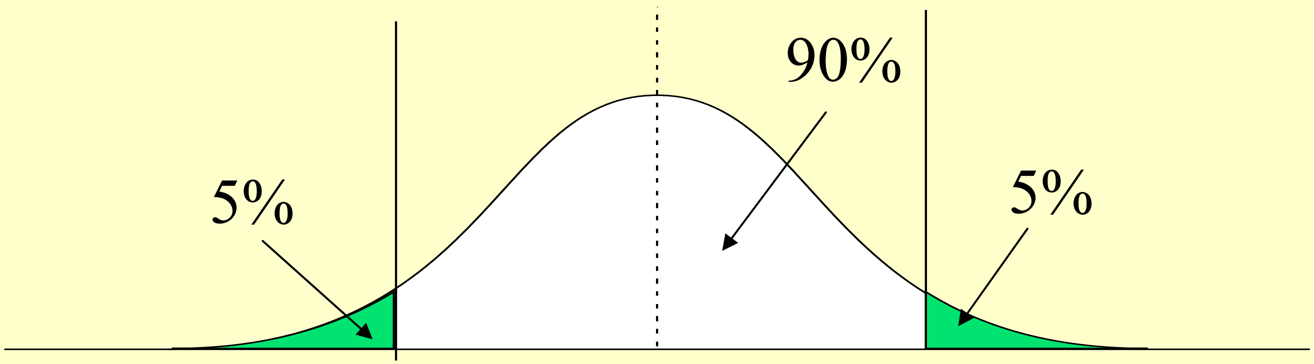
z	Height	Area
…	…	…
0.70	31.23	51.61
0.75	30.11	54.67
0.80	28.97	57.63
…	…	…

Another way to use the z-tables

- Mean SAT score = 500, std. deviation = 100
- Assuming that the distribution of scores is normal, what is the score such that 95% of the scores are below that value?



Using z-tables to find the 95 percentile point



- From the tables:

z	Height	Area
1.65	10.23	90.11

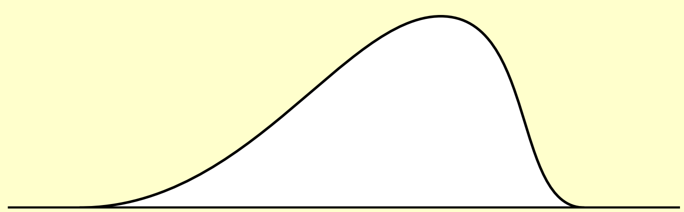
- z=1.65 -> x=? Mean=500, s.d.=100
- 1.65 = (x-500)/100; x = 165+500 = 665

Normal distributions

- A lot of data is normally distributed because of the central limit theorem from last time.
  - Data that are influenced by (i.e. the “sum” of) many small and unrelated random effects tend to be approximately normally distributed.
  - E.G. weight (I’m making up these numbers)
    - Overall average = 120 lbs for adult women
    - Women add about 1 lb/year after age 29
    - Illness subtracts an average of 5 lbs
    - Genetics can make you heavier or thinner
    - A given “sample” of weight is influenced by being an adult woman, age, health, genetics, …

Non-normal distributions

- For data that is approximately normally distributed, we can use the normal approximation to get useful information about percent of area under some fraction of the distribution.
- For non-normal data, what do we do?

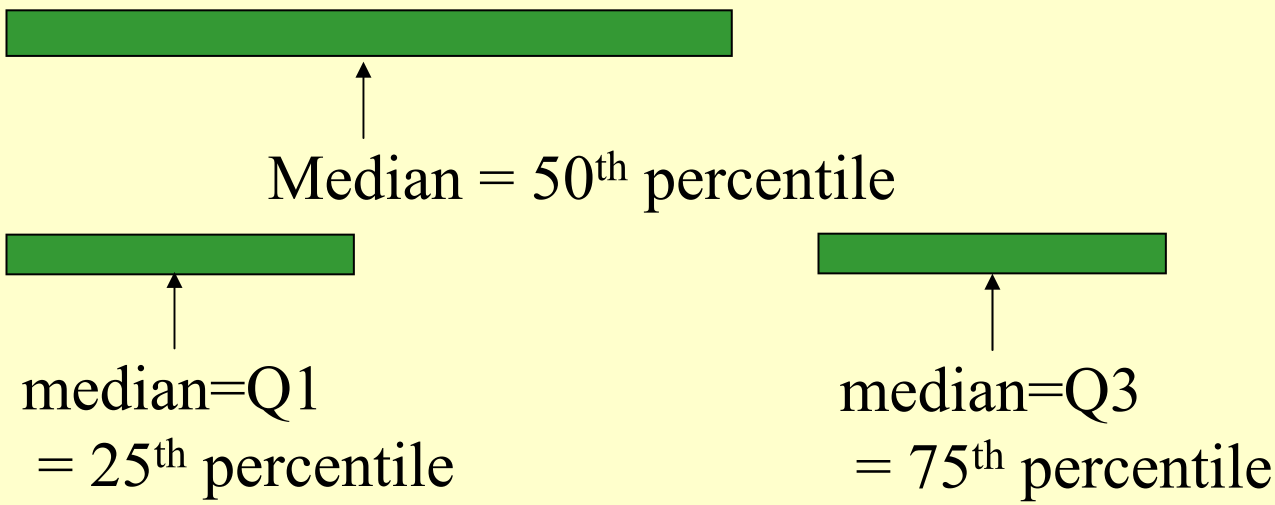


Non-normal distributions

- E.G. income distributions tend to be very skewed
- Can use percentiles, much like in the last z-table example (except without the tables)
  - What's the 10^{th} percentile point? The 25^{th} percentile point?

Percentiles & interquartile range

- Divide data into 4 groups, see how far about the extreme groups are.



- Q3-Q1 = IQR = 75^{th} percentile – 25^{th} percentile

What do you do for other percentiles?

- Median = point such that 50% of the data lies below that point
- Similarly, 10^{th} percentile = point such that 10% of the data lies below that point.

What do you do for other percentiles?

- If you have a theory for the distribution of the data, you can use that to find the nth percentile.
- Estimating it from the data, using MATLAB (to a first approximation)

  (x = the data)

  ```
  y = sort(x);
  N = length(x);      % how many data points there are
  TenthPerc = y(0.10*N);
  ```
- This isn’t exactly right (remember, for instance, that median (1 2 4 6) is 3), but it’s close enough for our purposes.

How do you judge if a distribution is normal?

- So far we've been eyeballing it. (Does it look symmetric? Is it about the right shape?) Can we do better than this?

Normal quantile Plots

- A useful way to judge whether or not a set of samples comes from a normal distribution.
- We'll still be eyeballing it, but with a more powerful visualization.

Normal quantile plots

data

↓

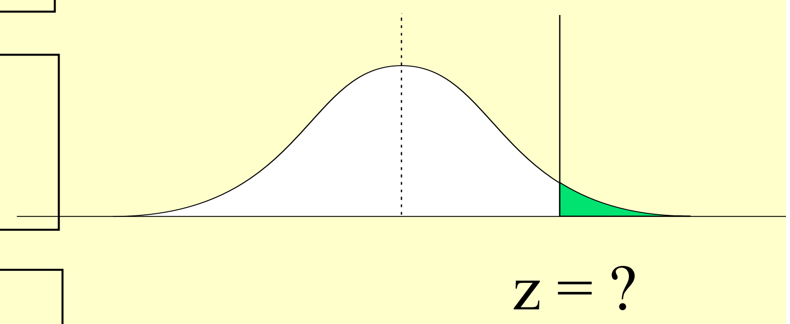
For each datum, what % of the data is below this value – what's its percentile?

↓

If this were a normal distribution, what z would correspond to that percentile?

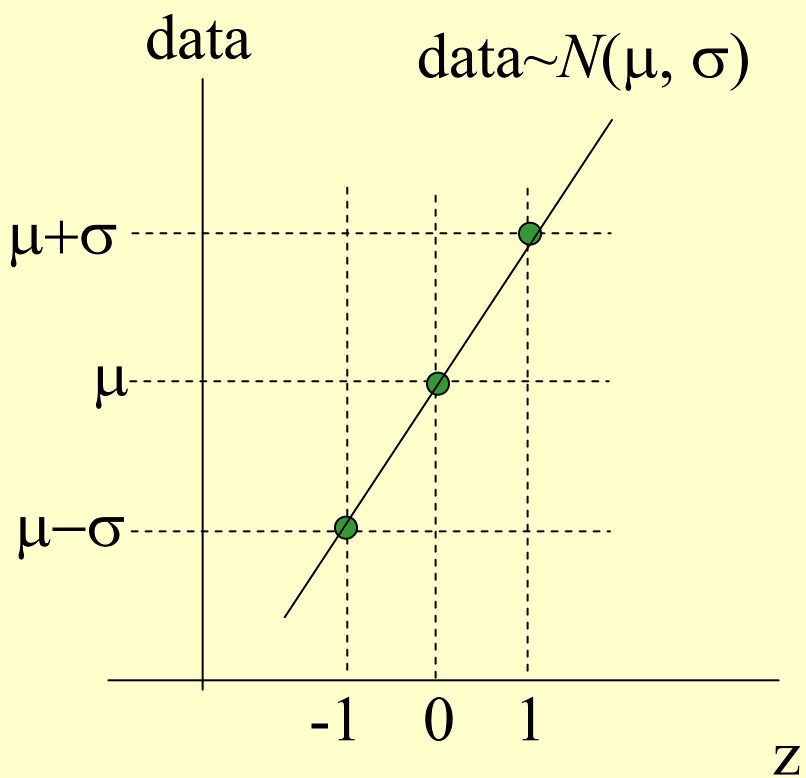
↓

Compare the actual data values to those predicted (from the percentiles) if it were a standard normal (z) distribution.



Normal quantile plots

- If the data~$N(0, 1)$, the points should fall on a 45 degree line through the origin.
- If the data~$N(\mu, 1)$, the points should fall on a 45 degree line.
- If the data~$N(\mu, \sigma)$, the points will fall on a line with slope σ (or $1/\sigma$, depending on how you plotted it).



Normal Quantile Plots

- Basic idea:
  - Order the samples from smallest to largest. Assume you have N samples. Renumber the ordered samples $\{x_1, x_2, \ldots, x_N\}$.
  - Each sample x_i has a corresponding percentile $k_i = (i-0.5)/N$. About k_i% of the data in the sample is $< x_i$.
  - If the distribution is normal, we can look up k_i % in the z-tables, and get a corresponding value for z_i.
  - Plot x_i vs. z_i (it doesn't matter which is on which axis)

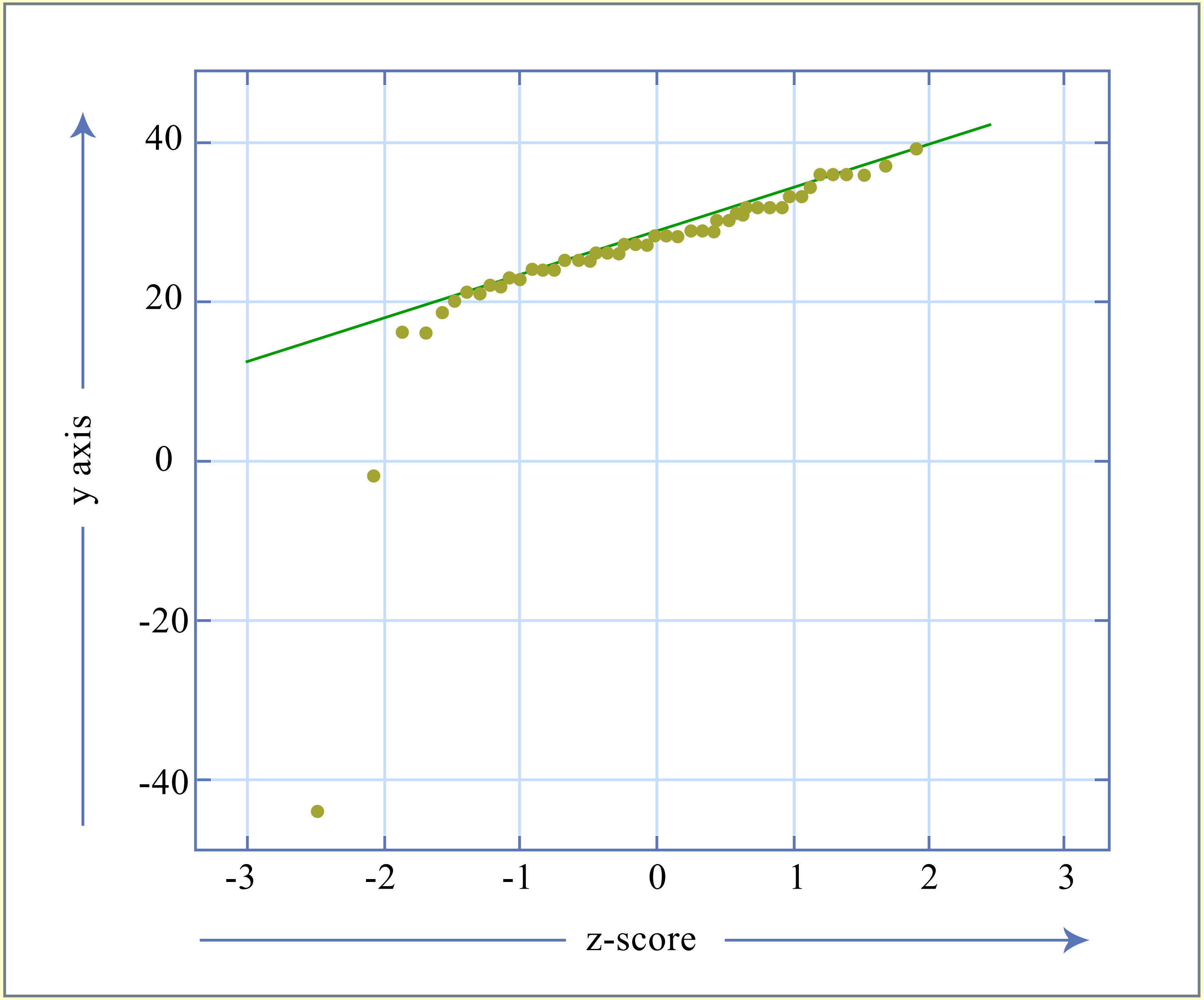


Figure by MIT OCW.

- Let’s remove those outliers…

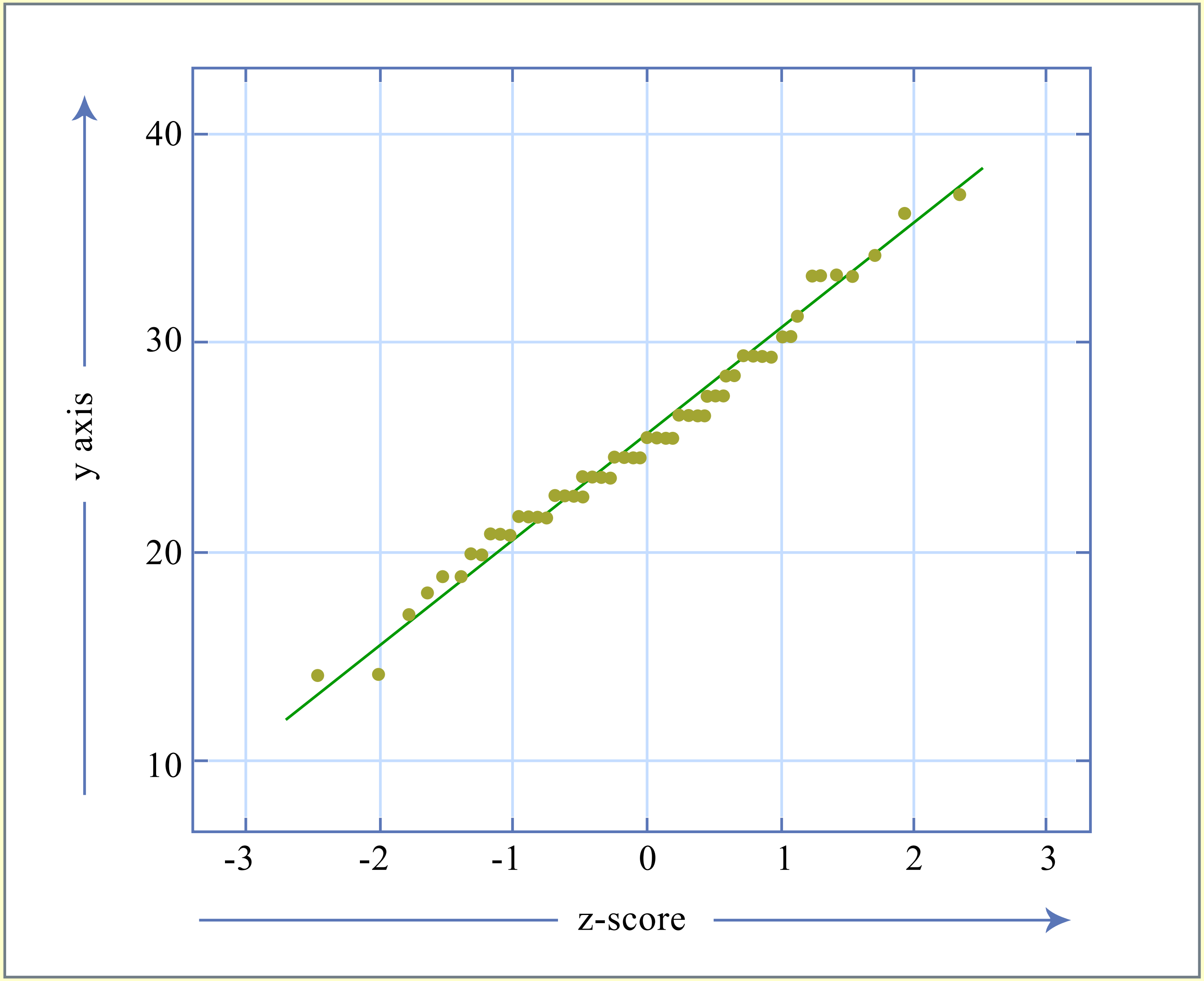
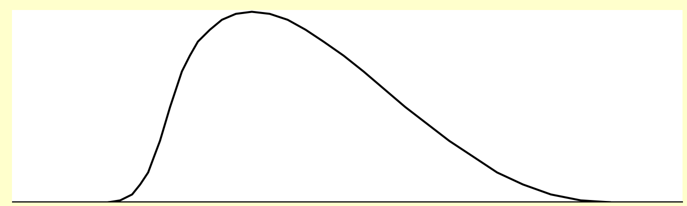


Figure by MIT OCW.

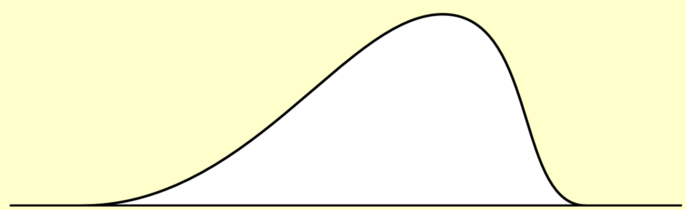
- The normal quantile plot allows us to see which points deviate strongly from a line. This helps us locate outliers.

Non-linear plots

- Concave-up (with the axes as shown here) means positive skew



- Concave-down means negative skew



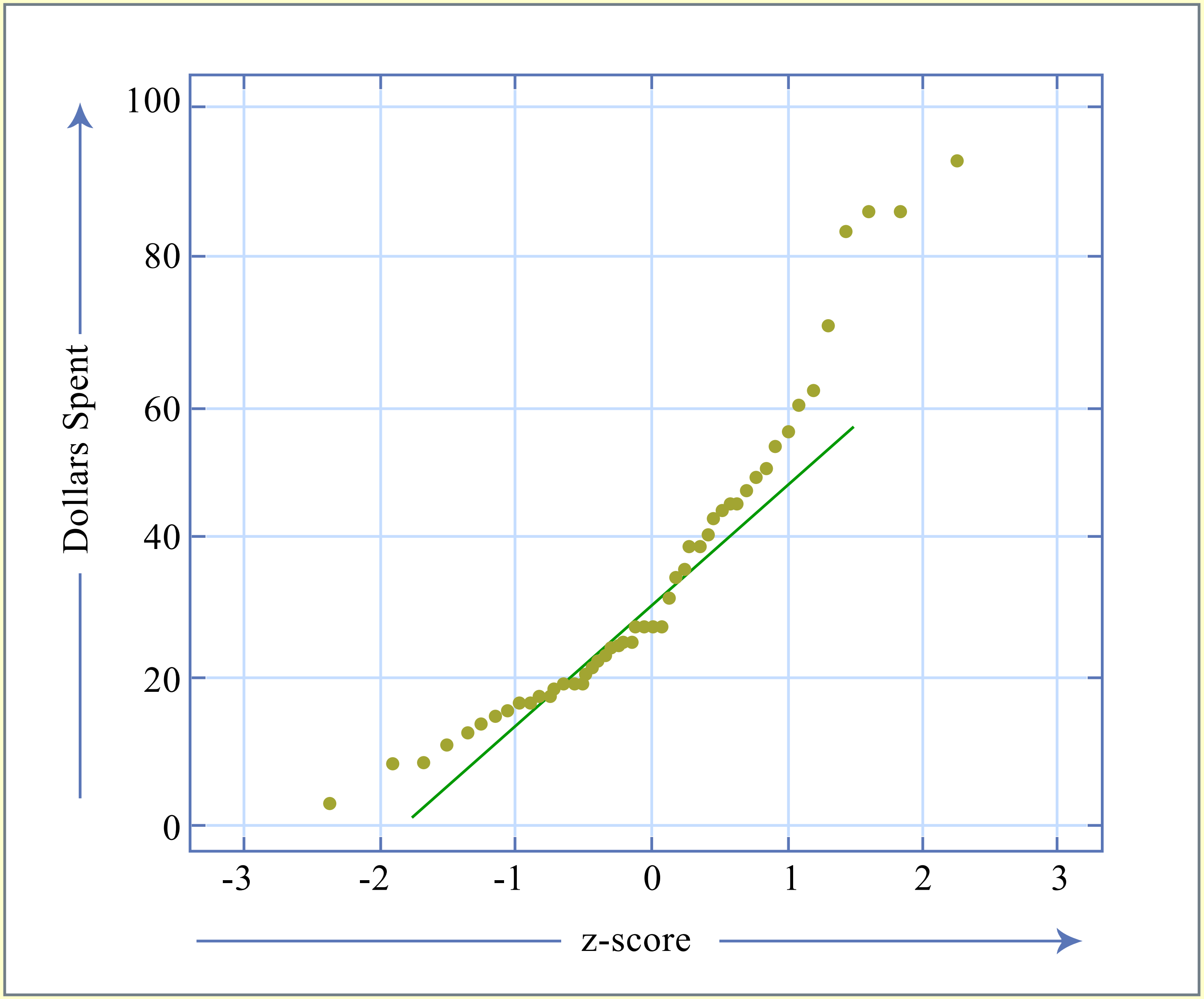


Figure by MIT OCW.

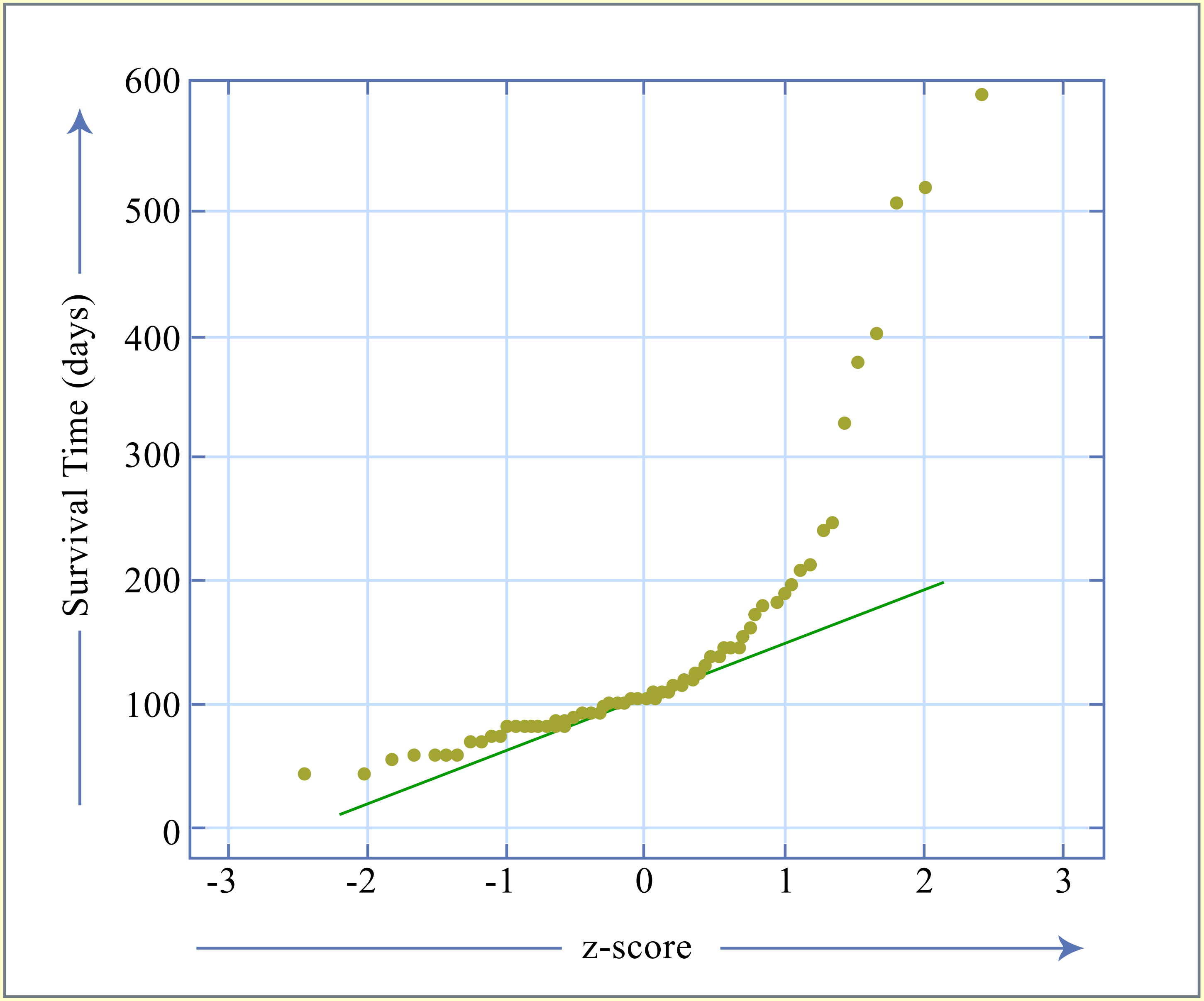


Figure by MIT OCW.

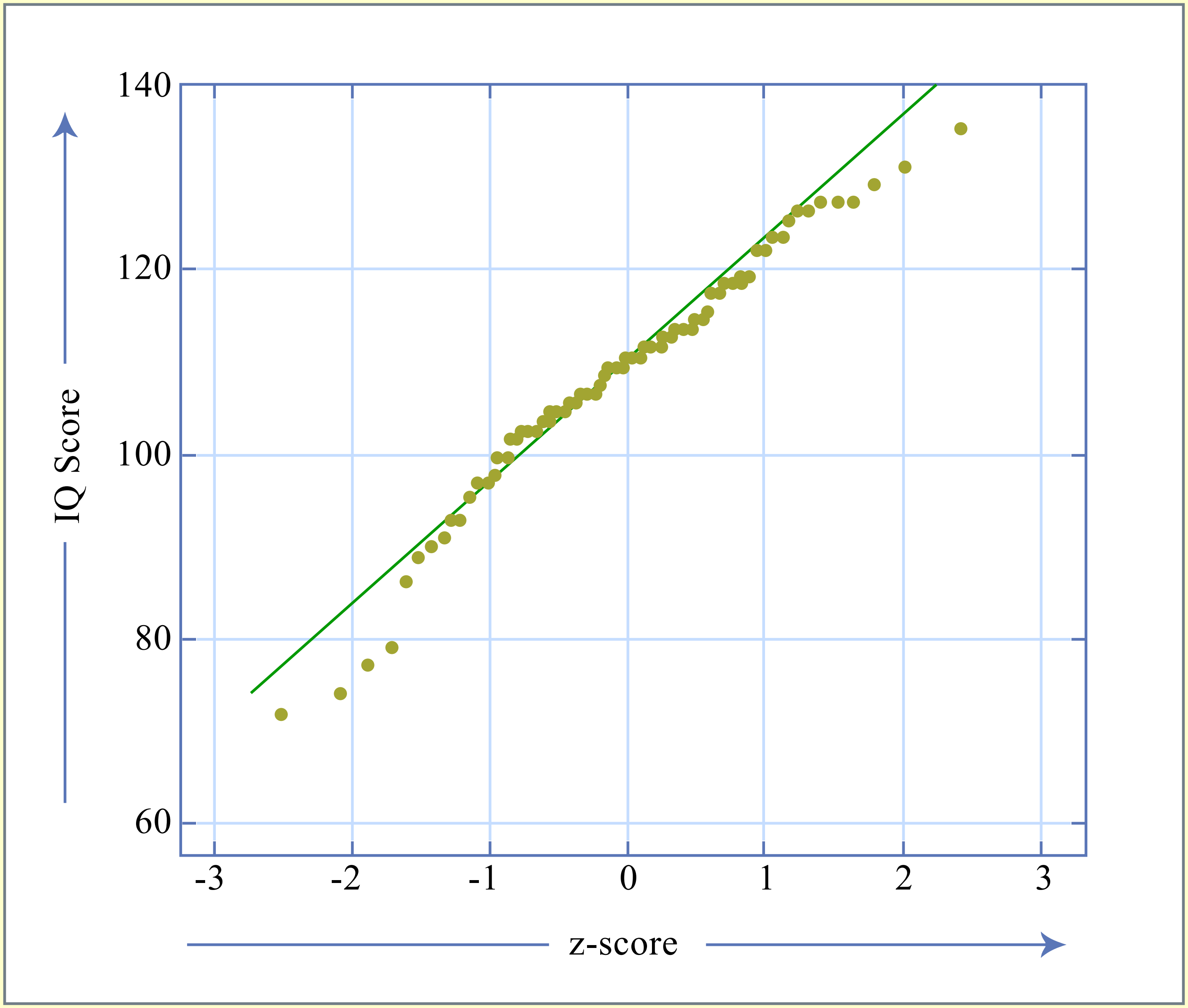


Figure by MIT OCW.

- Granularity
  - When the r.v. can only take on certain values, the normal quantile plot looks like funny stair steps
  - E.G. binomial distributions – we'll get there in a sec.

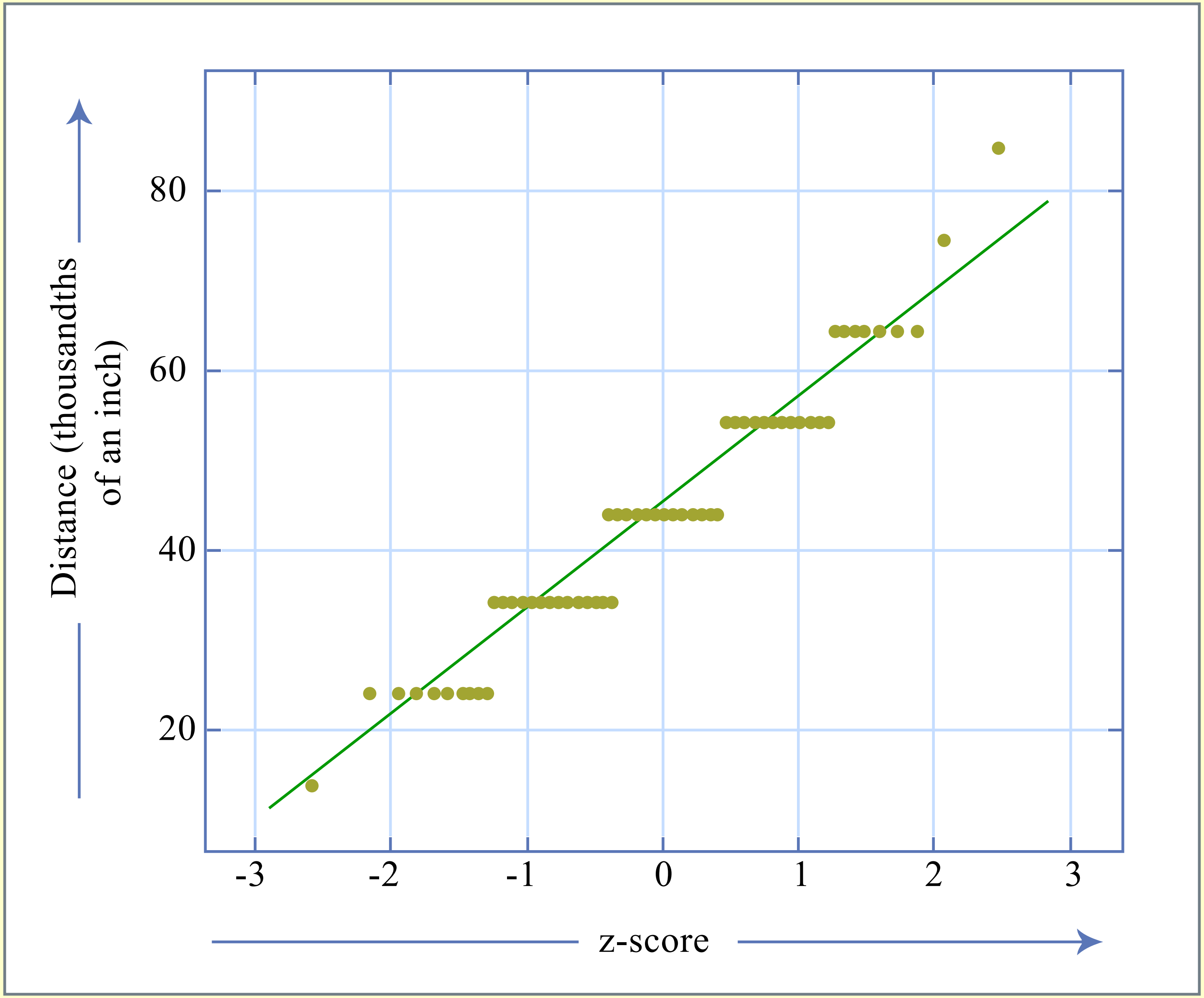


Figure by MIT OCW.

Normal quantile plots in MATLAB

- qqplot(x) generates a normal quantile plot for the samples in vector x
- You should have access to this command on the MIT server computers.

The binomial distribution

- An important special case of a probability distribution.
- One of the most frequently encountered distributions in statistics
- Two possible outcomes on each trial, e.g. {H, T}
- One outcome is designated a "success", the other a "failure"
- The binomial distribution is the distribution of the number of successes on N trials.
- E.G. the distribution of the number of heads, when you flip the coin 10 times.

Example

- Flip a fair coin 6 times.
- What is P(4H, 2T)?
- Well, first, note that P(TTHHHH) = P(THHHHT) = … = $(0.5)^4 (1-0.5)^2 = (0.5)^6$
  - All events with 4H have the same probability
  - How many such events are there?
- P(4H, 2T) =

  (# events of this type) x $(0.5)^4 (1-0.5)^2$

How many events of this type are there? The binomial coefficient

- Equals number of possible combinations of N draws such that you have k successes.

$$\equiv \binom{N}{k} = \frac{N!}{k!(N-k)!}$$

- N! = N factorial = N(N-1)(N-2)…(1) = factorial(N) in MATLAB
- 0! = 1

Intuition for the binomial coefficient

- N! = number of possible ways to arrange 6 unique items (a,b,c,d,e,f)
  - 6 in 1st slot, 5 remain for 2nd slot, etc.
- But, they aren't unique. k are the same (successes), and the remaining (N-k) are the same (failures).
- k! and (N-k)! are the # of "duplicates" you get from having k and N-k items be the same.
- The result is the number of combinations with k successes.

Binomial coefficient

- Number of ways of getting k heads in N tosses
- Number of ways of drawing 2 R balls out of 5 draws, with p(R) = 0.1
- Number of ways of picking 2 people out of a group of 5 (less obvious)
  - Associate an indicator function with each person = 1 if picked, 0 if not
  - p(p1 = 1) is like p(toss 1 = H)

The Binomial distribution

- Probability of k successes in N tries
- Repeatable sampling of a binomial variable (e.g., tossing a coin), where you *decide the number of samples in advance*
  - (versus: I keep drawing a ball until I get 2 reds, then I quit. What was my probability of getting 2R and 3G?)
- Three critical properties
  - Result of each trial may be either a failure or a success
  - Probability of success is the same for each trial
  - The trials are independent

Back to tossing coins…

- The coin-toss experiment is an example of a binomial process
- Let’s arbitrarily designate “heads” as a success
- p(heads) = 0.5
- What is the probability of obtaining 4 heads in 6 tosses?

Example

- P(4 H in 6 tosses) =

$$\binom{N}{k} p^k (1-p)^{N-k}$$

$$= \binom{6}{4} (0.5)^4 (0.5)^2$$

$$= \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1 \cdot 2 \cdot 1} \times \frac{1}{64} = \frac{15}{64}$$

Kangaroo example from book

- 10 pairs of kangaroos
- Half of them get vitamins
- 10 races (vitamin vs. no vitamin)
- 7 out of 10 races, the kangaroo taking vitamins wins
- Do the vitamins help, or is this just happening by chance?

How do we decide?

- What we want to do is to set a criterion # of wins, and decide that the vitamins had an effect if we see a # of wins equal to or greater than the criterion.
- How do we set the criterion?
- Well, what if we had set the criterion right at 7 wins? What would be our probability of saying there was an effect of the vitamins, when really the results were just due to chance?

Roo races

- If we set the criterion at 7 wins, and there were no effect of vitamins, what is the probability of us thinking there were an effect?
- Probability of the vitamin roo winning, if vitamins don’t matter, = p = 0.5
- What is the probability, in this case, of 7 wins, or 8, or 9, or 10?

Roo races

- P(7 wins out of 10) + P(8 wins out of 10) + P(9 wins out of 10) + P(10 wins out of 10)
- Use the binomial formula, from before.
- $\approx$ 17% (see problem 6, p. 258, answer on p. A-71)

Roo races

- Remember, this is the probability of us thinking there were an effect, when there actually wasn’t, if we set the criterion at 7 wins.
- 17% is a pretty big probability of error. (In statistics we like numbers more like 5%, 1%, or 0.1%.)
- We probably wouldn’t want to set the criterion at 7 wins. Maybe 8 or 9 would be better.
- We decide that the vitamins probably have no effect.

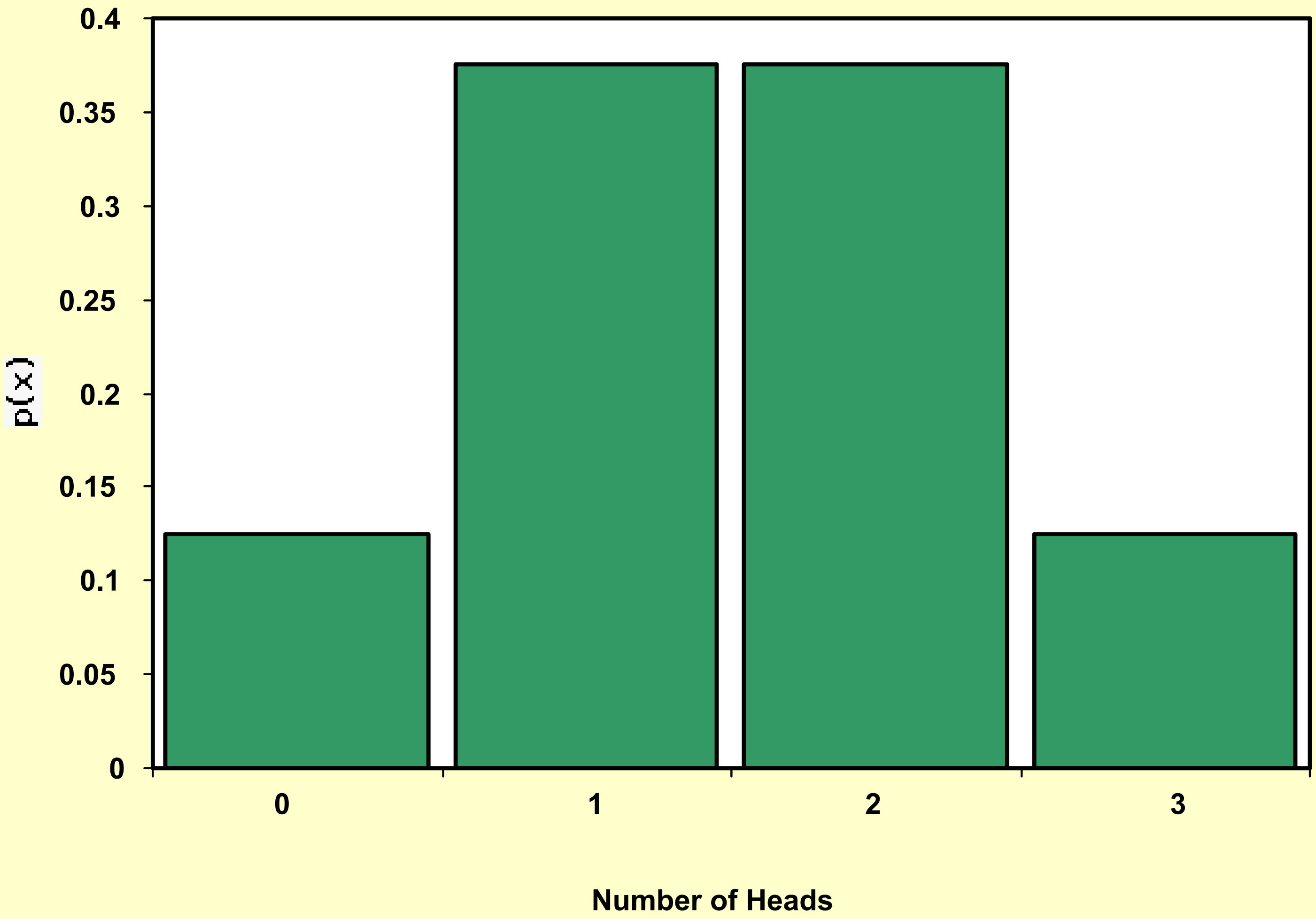
- We’ll see LOTS more problems like the kangaroo problem in this class.
- And this whole business of setting a criterion will become more familiar and intuitive.
- For now, back to binomial random variables.

Mean and variance of a binomial random variable

- The mean number of successes in a binomial experiment is given by:
  - $\mu = np$
  - n is the number of trials, p is the probability of success
- The variance is given by
  - $\sigma^2 = npq$
  - $q = 1-p$

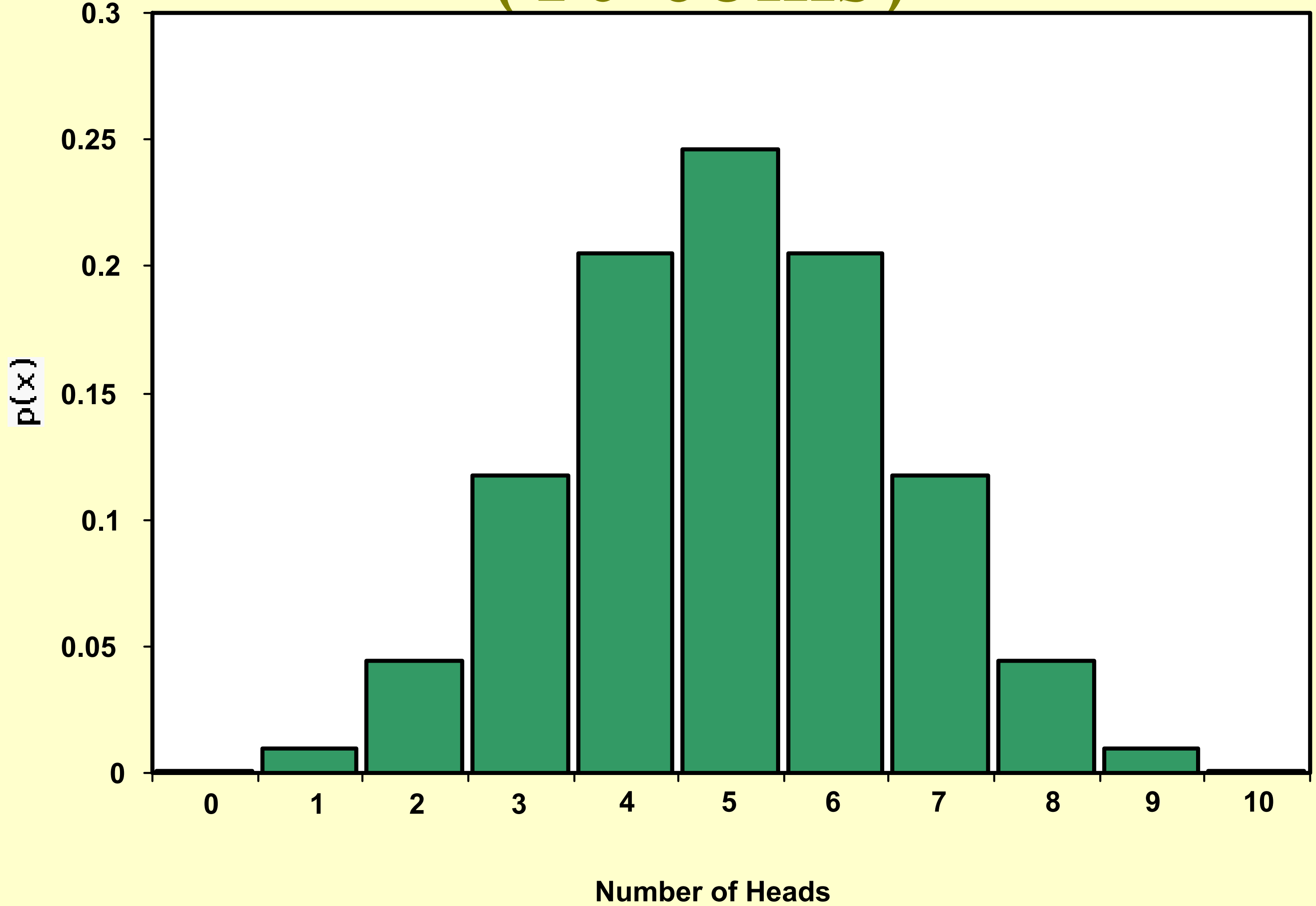
What happens to the binomial distribution as you toss the coin more times?

Probability Histogram (3 coins)

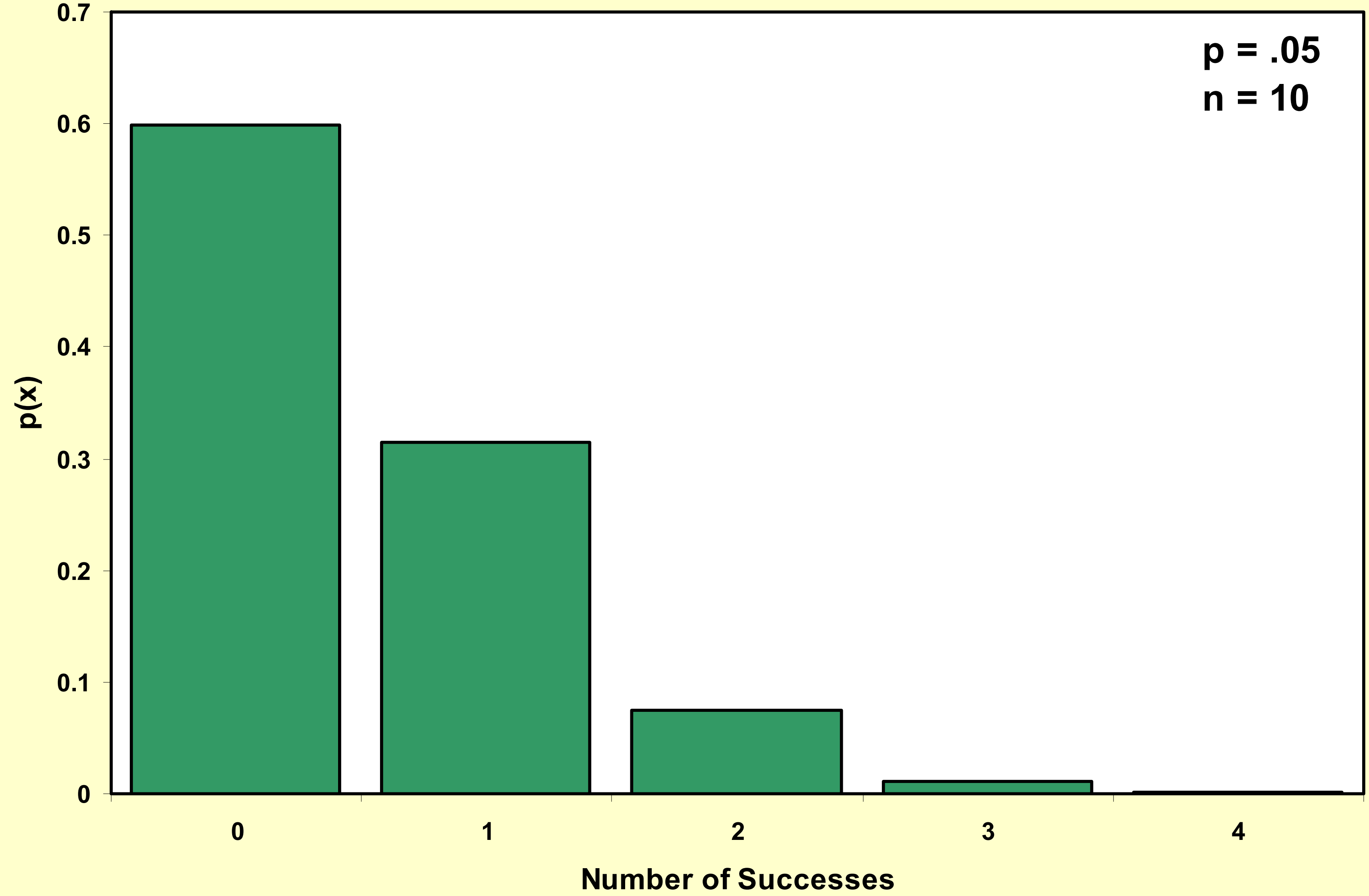


Probability Histogram (4 coins)

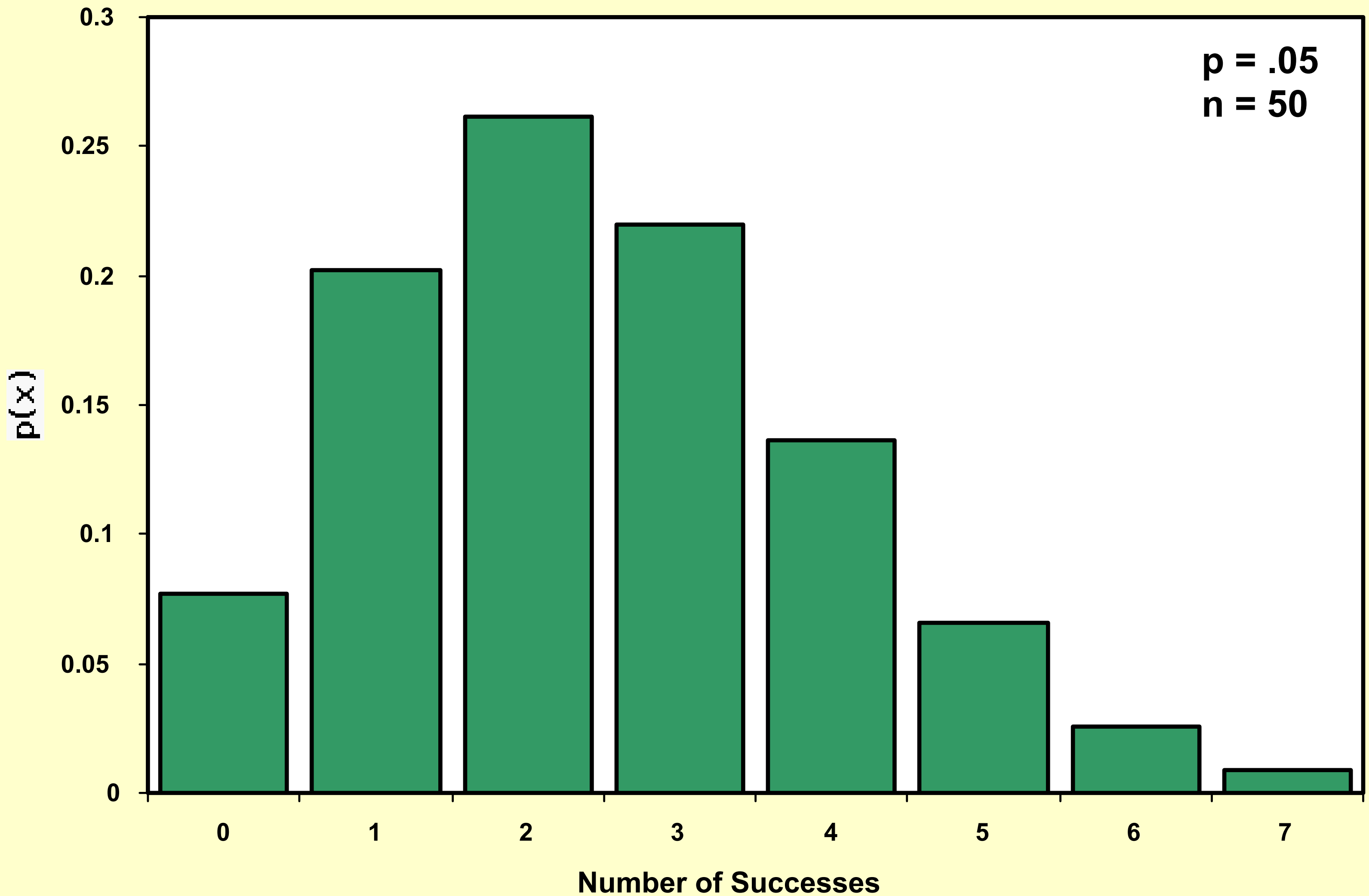
Probability Histogram (10 coins)



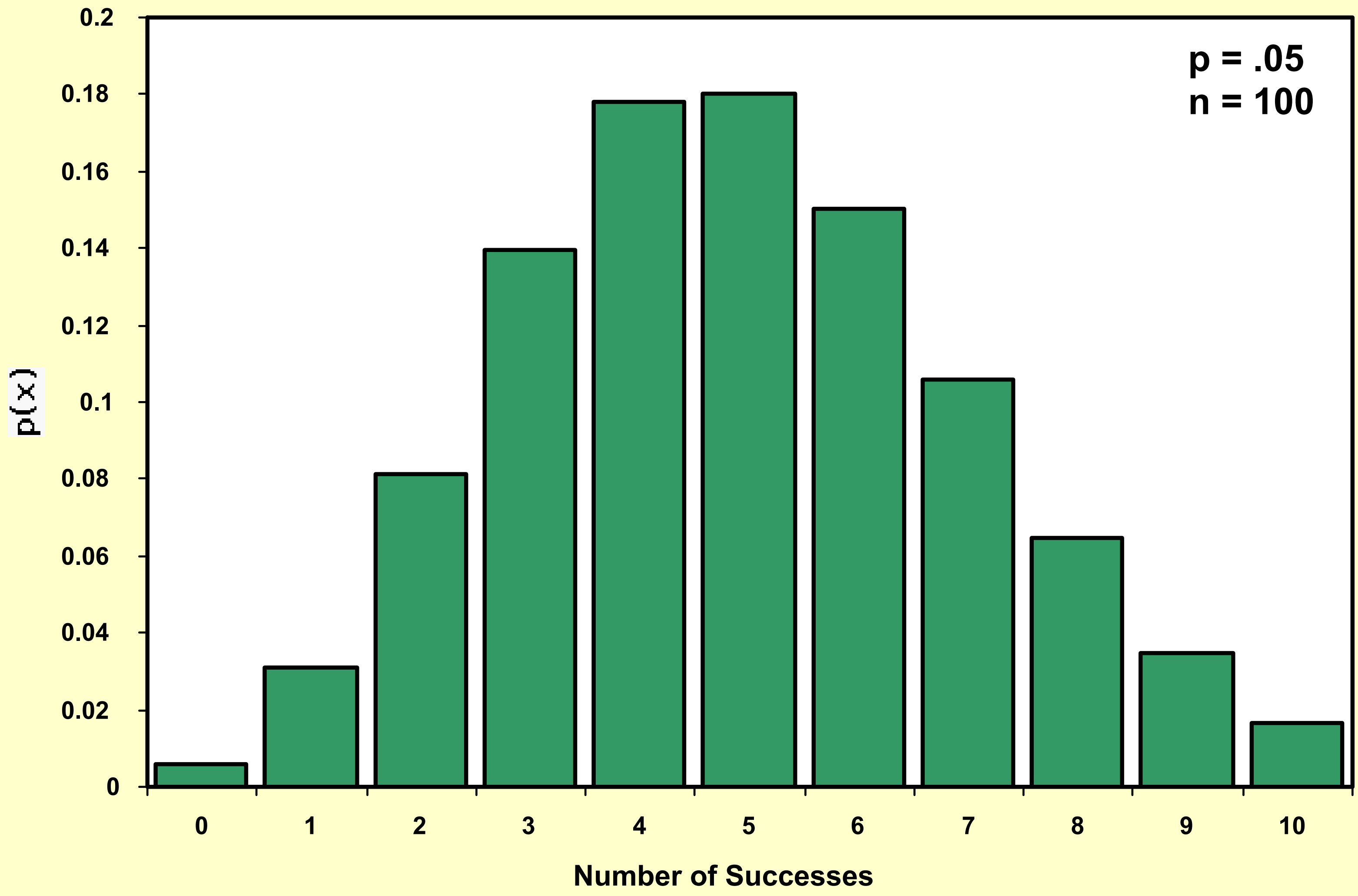
Binomial Distribution



Binomial Distribution



Binomial Distribution



The central limit theorem, again

- As the number of tosses goes up, the binomial distribution approximates a normal distribution.
- The total number of heads on 100 coin tosses = number on 5 tosses + number on next 5 tosses + …
- Thus, a binomial process can be thought of as the sum of a bunch of independent processes, the central limit theorem applies, and the distribution approaches normal, for a large number of “coin tosses” = trials.

The normal approximation

- This means we can use z-tables to answer questions about binomial distributions!

Normal Approximation

- When is it OK to use the normal approximation?

- Use when n is large and p isn't too far from 0.5
  - The further p is from .5, the larger n you need
  - Rule of thumb: use when $np \geq 10$ and $nq \geq 10$

Normal Approximation

- For any value of p, the binomial distribution of n trials with probability p is approximated by the normal curve with
  - $\mu = np$ and
  - $\sigma = sqrt(npq)$
    - Where $q = (1\text{-}p)$
- Let's try it for 25 coin flips...

25 coin flips

- What is the probability that the number of heads is ≤ 14?
- We can calculate from the binomial formula that $p(x \leq 14)$ is .7878 exactly

Normal Approximation

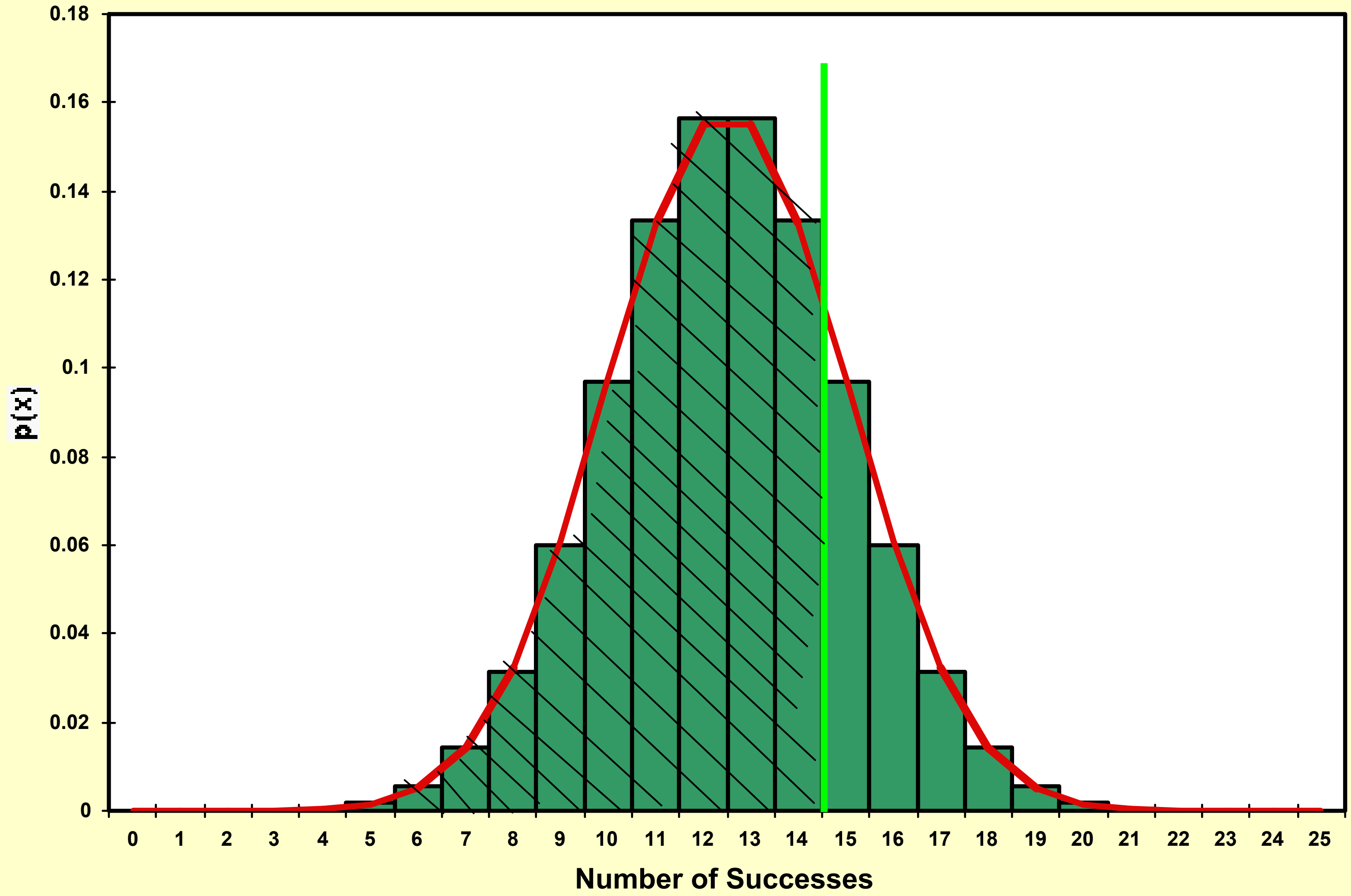
- Using the normal approximation with

  $\mu = np = (25)(5) = 12.5$ and

  $\sigma = \text{sqrt}(npq) = \text{sqrt}((25)(.5)(.5)) = 2.5$ we get
- $p(x \leq 14) = p(z \leq (14\text{-}12.5)/2.5))$

  $= p(z \leq .6) = .7257$
- .7878 vs. .7257 -- not great!!
- Need a better approximation...

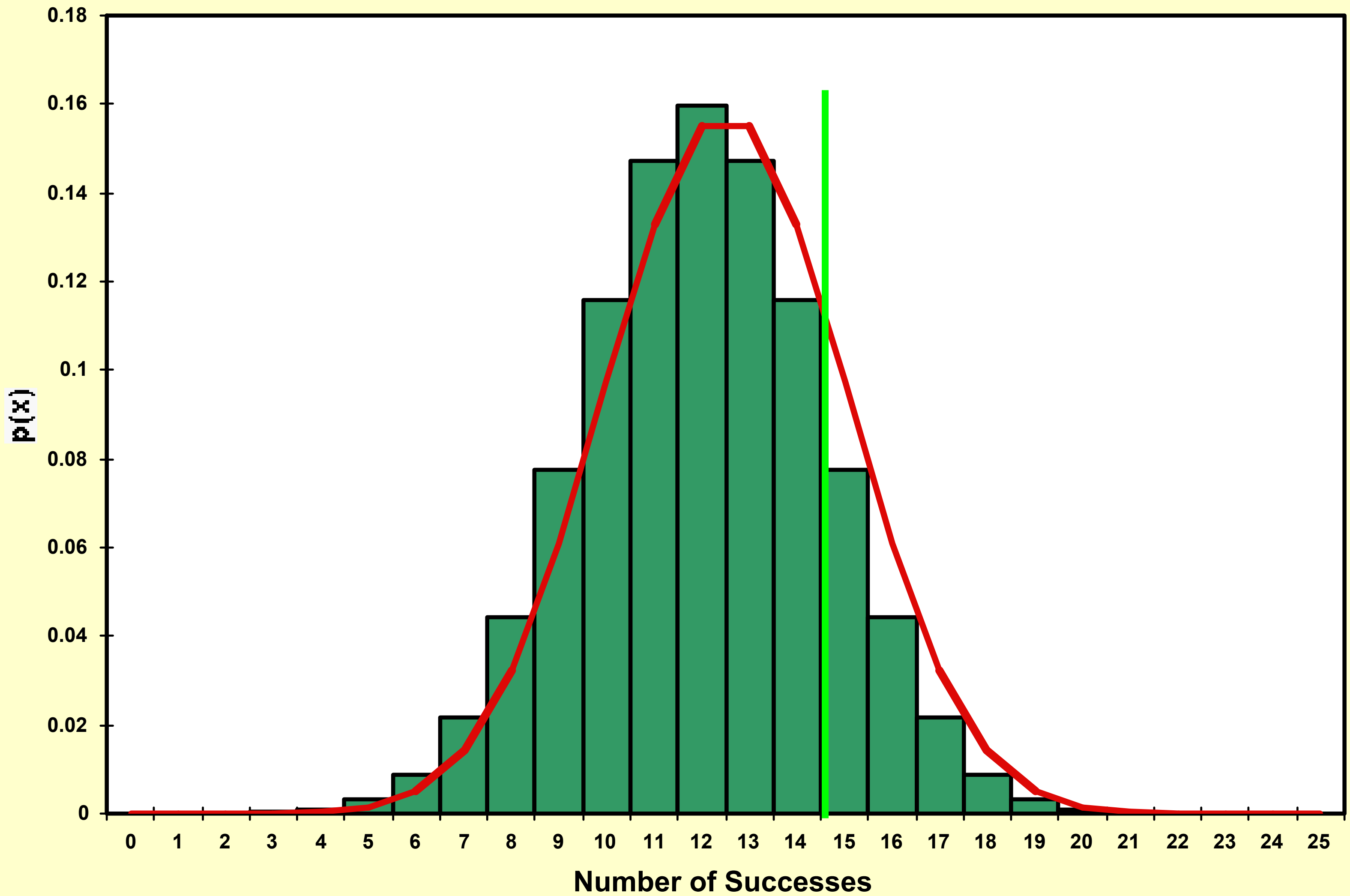
Normal Approximation of Binomial Distribution



Continuity Correction

- Notice that the bars are centered on the numbers
- This means that $p(x \leq 14)$ is actually the area under the bars less than $x=14.5$
- We need to account for the extra 0.5
- $P(x \leq 14.5) = p(z \leq .8) = .7881$ -- a much better approximation!

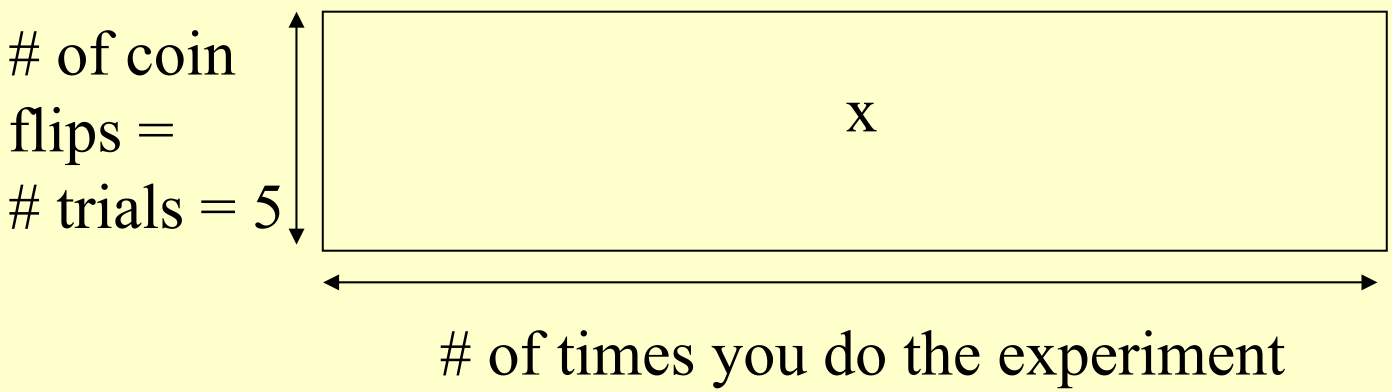
Continuity Correction

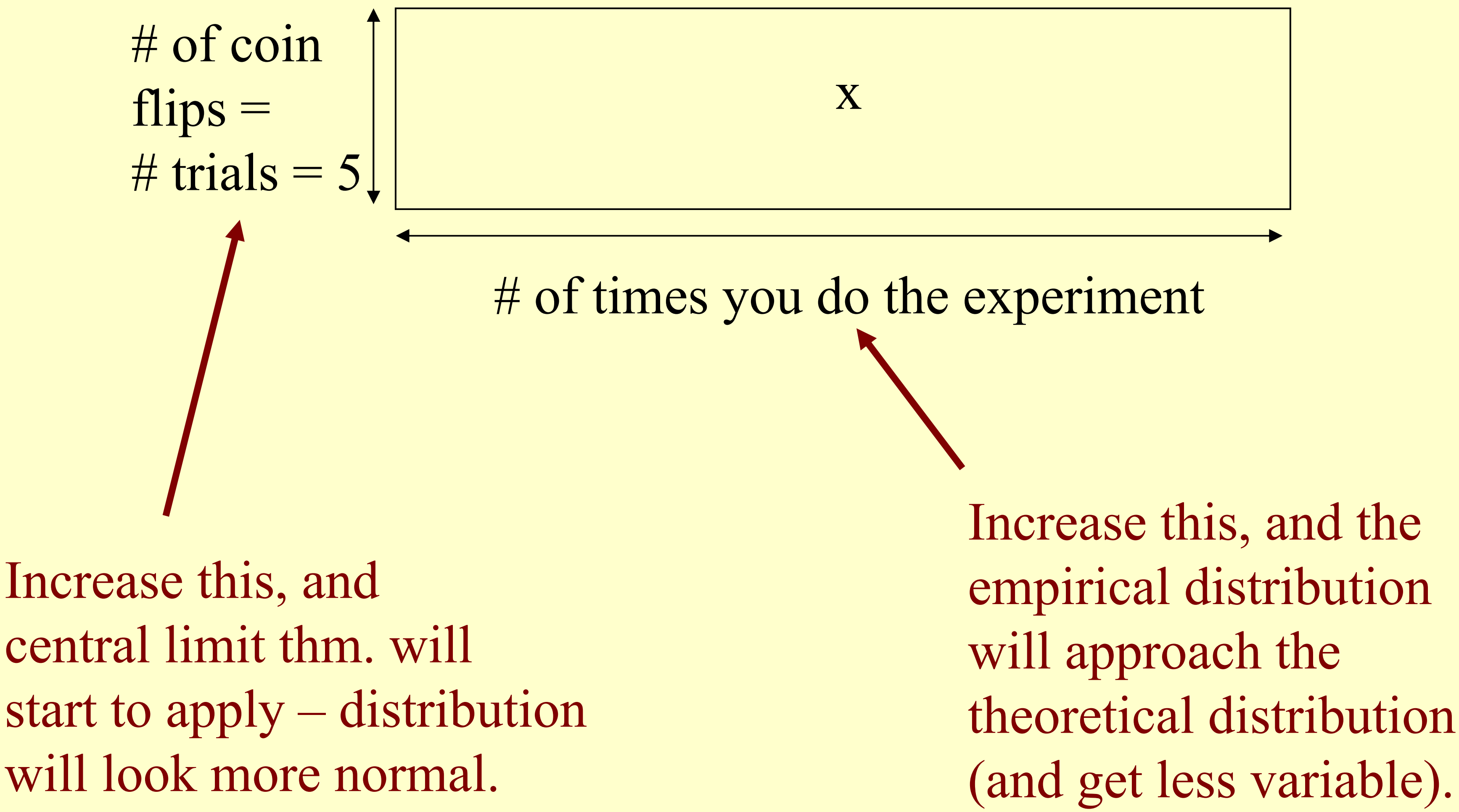


of times you do an experiment, vs. # of trials in that experiment

- In MATLAB:

```
x = rand(5,10000);
coinflip = x>0.5;        % 1 = heads
y = sum(x);      % number of heads
```





# of coin
flips =
trials = 5
x
# of times you do the experiment
Increase this, and
central limit thm. will
start to apply – distribution
will look more normal.
Increase this, and the
empirical distribution
will approach the
theoretical distribution
(and get less variable).

Binomial distribution and percent

- Can also use binomial distribution for percent "success", by dividing by the number of samples (trials)
- Mean = np/n = p
- Std. deviation = sqrt(npq)/n = sqrt(pq/n)
- We'll use this a lot in class, as we often have a situation like that for elections: 45% favor Kerry, 39% favor Edwards – are these different by chance, or is there a real effect there?