

Data Analysis and Statistical Methods

Statistics 651

<http://www.stat.tamu.edu/~suhasini/teaching.html>

Lecture 8 (MWF) The binomial and introducing the normal distribution

Suhasini Subba Rao

From binomial to normal

- In this lecture we introduce the normal distribution.
- It is commonly used in statistics. We could introduce as a distribution in its own right. But it is instrumental in statistics. Thus we motivate its existence through the binomial distribution (which pre-dates the normal distribution).
- We show that binomial distribution “converges” for a large sample size. This means the shape of the binomial looks more and more like the normal as the sample size grows.
- This is because the binomial distribution is the sum of binary random variables, where $S_n = X_1 + \dots + X_n$, this is the number of successes in n “trials”.

- Through a series of plots, we will illustrate the effect of the central limit theorem.

Reminder: The z-score (Lecture 4, slide 21)

- To measure the relative distance of an observation from its mean we calculate the z-score.
- For example, suppose the mean iron level in a blood sample is 20. The reading of patient of being 14 means nothing. All we can say is that it six less than the mean, but we know nothing else. But the standard deviation gives extra information. If the standard deviation of the blood sample is 8, then the relative distance is

$$z = \frac{\text{observation} - \text{mean}}{\text{s.d}} = \frac{14 - 20}{8} = -\frac{3}{4}.$$

The z-score tells us how close the observation to a “normal” reading. The z-score is an important transformation in statistics.

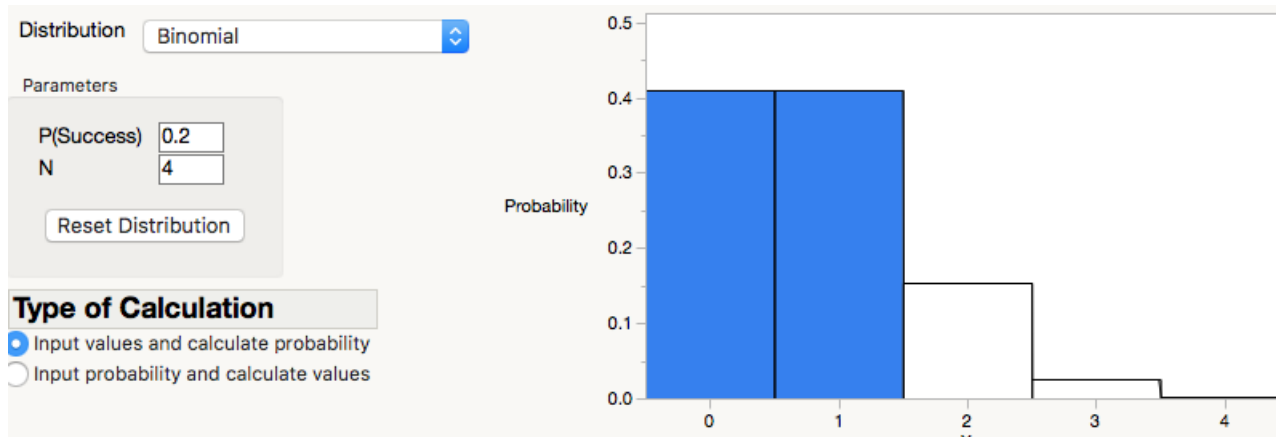
The mean and standard deviation of a binomial

Example 1: $n=4$ and $p=0.2$ Probabilities and plot of S_4

$$P(S_4 = 0) = (0.8)^4, \quad P(S_4 = 1) = 4 \times (0.8)^3 \times (0.2)$$

$$P(S_4 = 2) = 6 \times (0.8)^2 \times (0.2)^2, \quad P(S_4 = 3) = 4 \times (0.8)^1 \times (0.2)^3$$

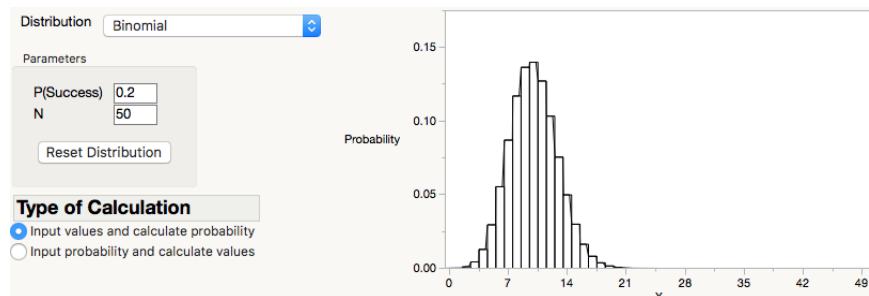
$$P(S_4 = 4) = (0.2)^4$$



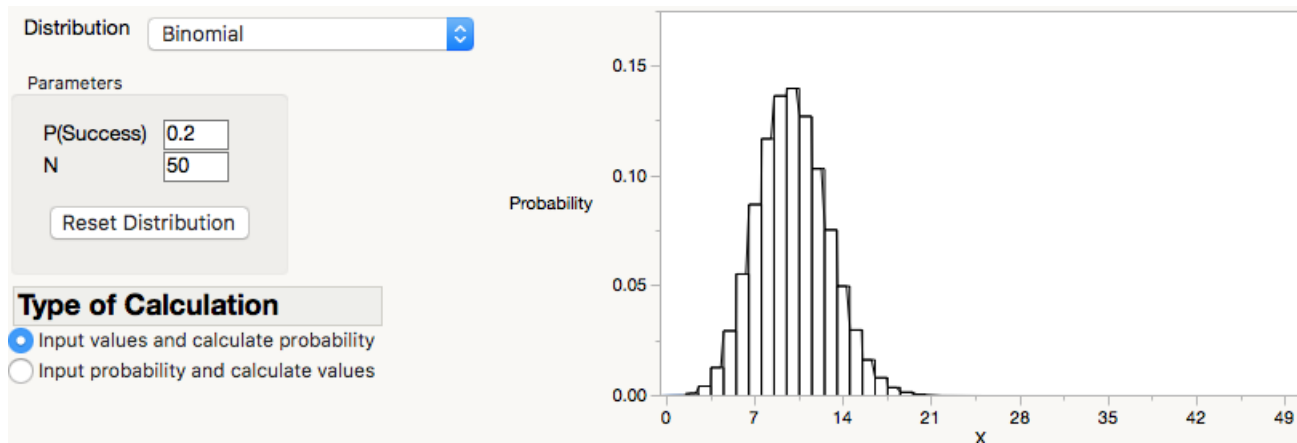
- The plot must have a center and spread, therefore it has a mean and a variance.
- The mean is 4×0.2 . This makes sense, in 4 trials the average number of successes will be 20% of 4.
- The corresponding standard deviation $\sqrt{4 \times 0.2 \times (1 - 0.2)}$. Remember this is the “average distance of S_n ” from the mean 4.

Example 2: $n=50$ and $p=0.2$

- Suppose there are 50 exam questions. There is 20% getting one correct. The score is S_n and $S_n \sim \text{Bin}(50, 0.2)$.
- The average number of questions one will correctly answer is 20% of 50 $= 0.2 \times 50 = 10$ (formula mean $= n \times p$).
- The variation is measured with a standard deviation. The variation for more score (or standard deviation) is $s = \sqrt{50 \times 0.2 \times 0.8} = \sqrt{8}$.



- The distribution is not exactly symmetric. But one can say at around 10 it is close to symmetric.

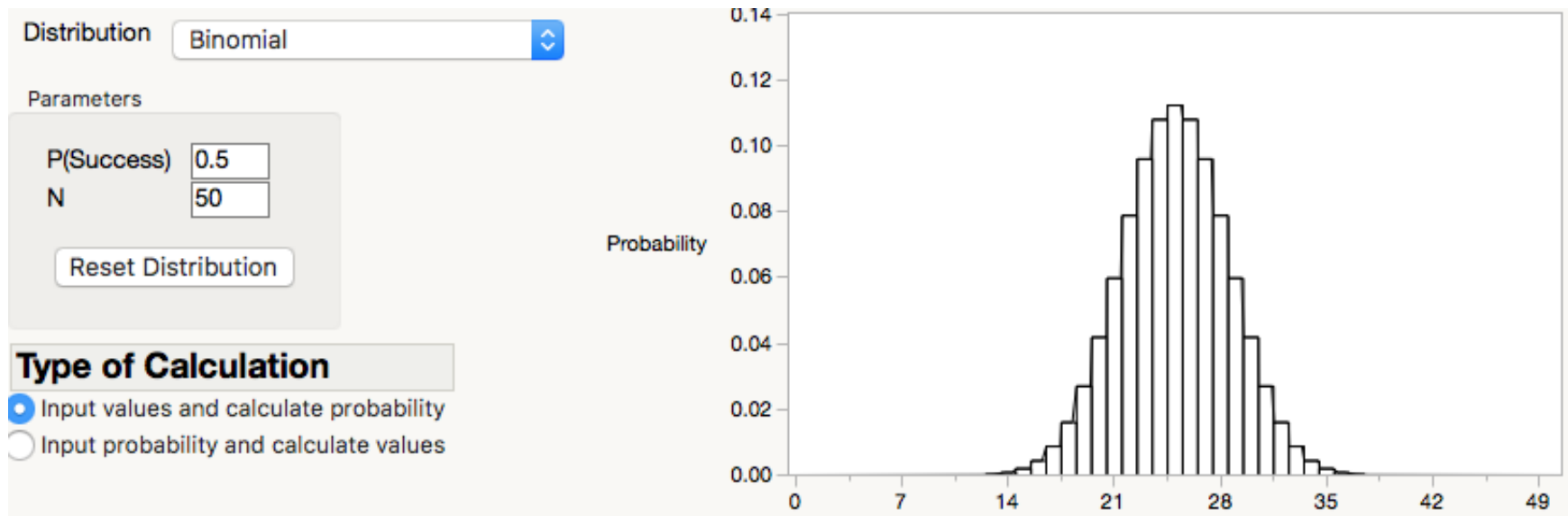


- The only difference between this example and the previous example is that $n = 4$ changed to $n = 50$.

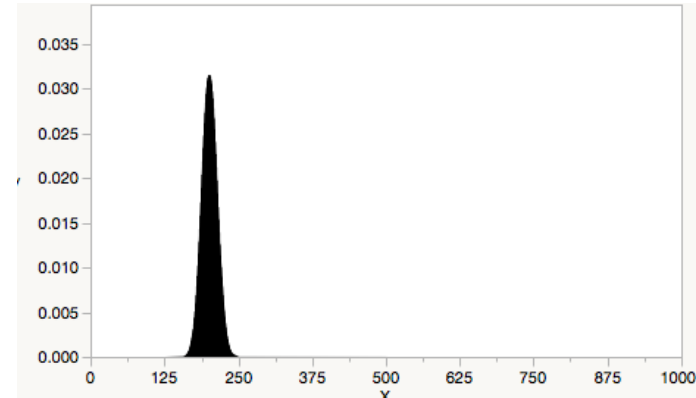
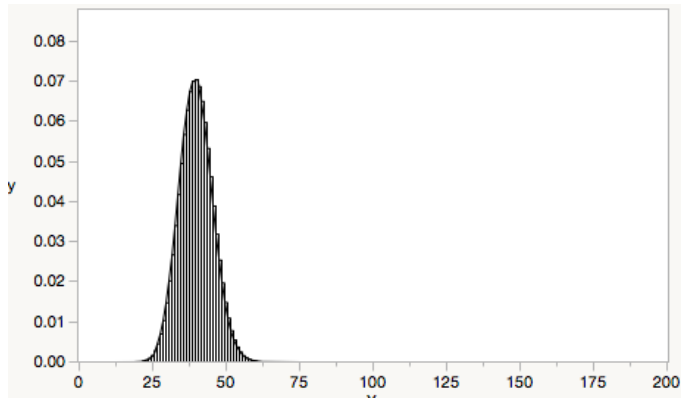
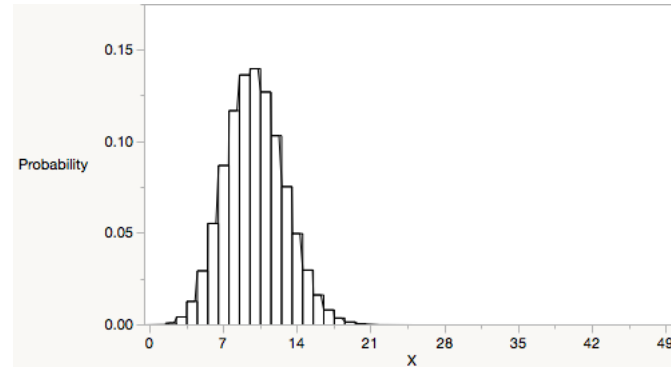
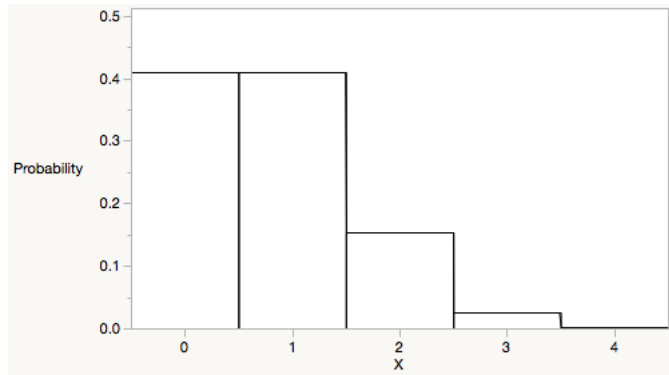
Example 3: $n=50$ and $p=0.5$

- Suppose now the exam is a true or false exam. The probability of getting a question correct by randomly guessing is 50%. There are 50 questions on the exam. If the student only guesses, their grade is random and follows a $\text{Bin}(50, 0.5)$.
- We are interested in the score out of 50, which is $S_{50} \sim \text{Bin}(50, 0.5)$ (the number of successes out of 50).
 - If the student guesses the average grade is $50 \times 0.5 = 25$.
 - The standard deviation of the grade is standard deviation $\sqrt{50 \times 0.5 \times 0.5} = \sqrt{12.5} \approx 3.5$.

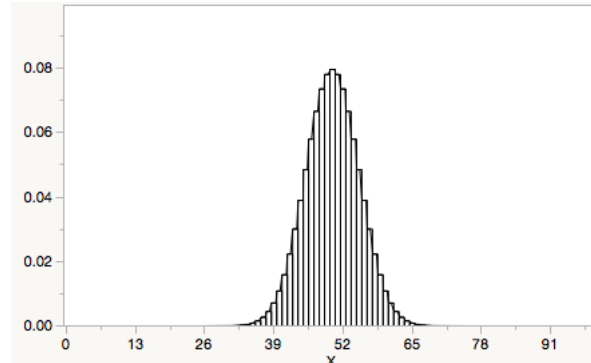
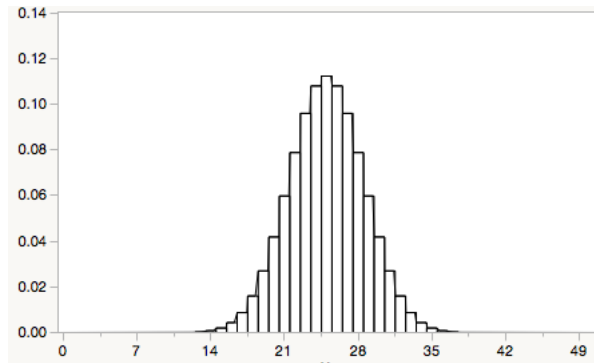
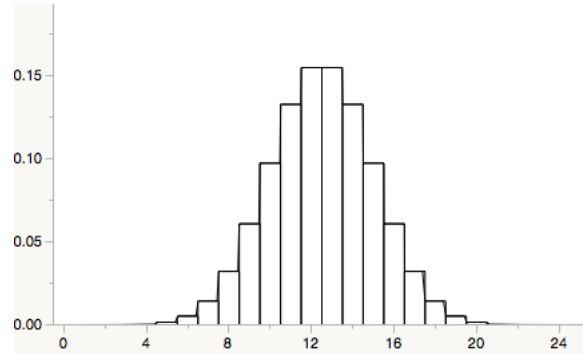
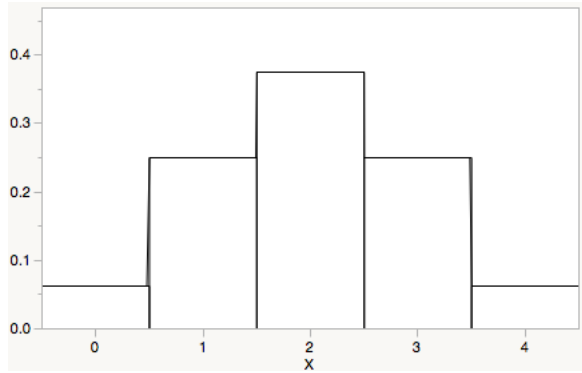
- The distribution is symmetric about the mean, 25. This is because $p = 0.5$, so there is an equal chance of the score being above or below the mean.



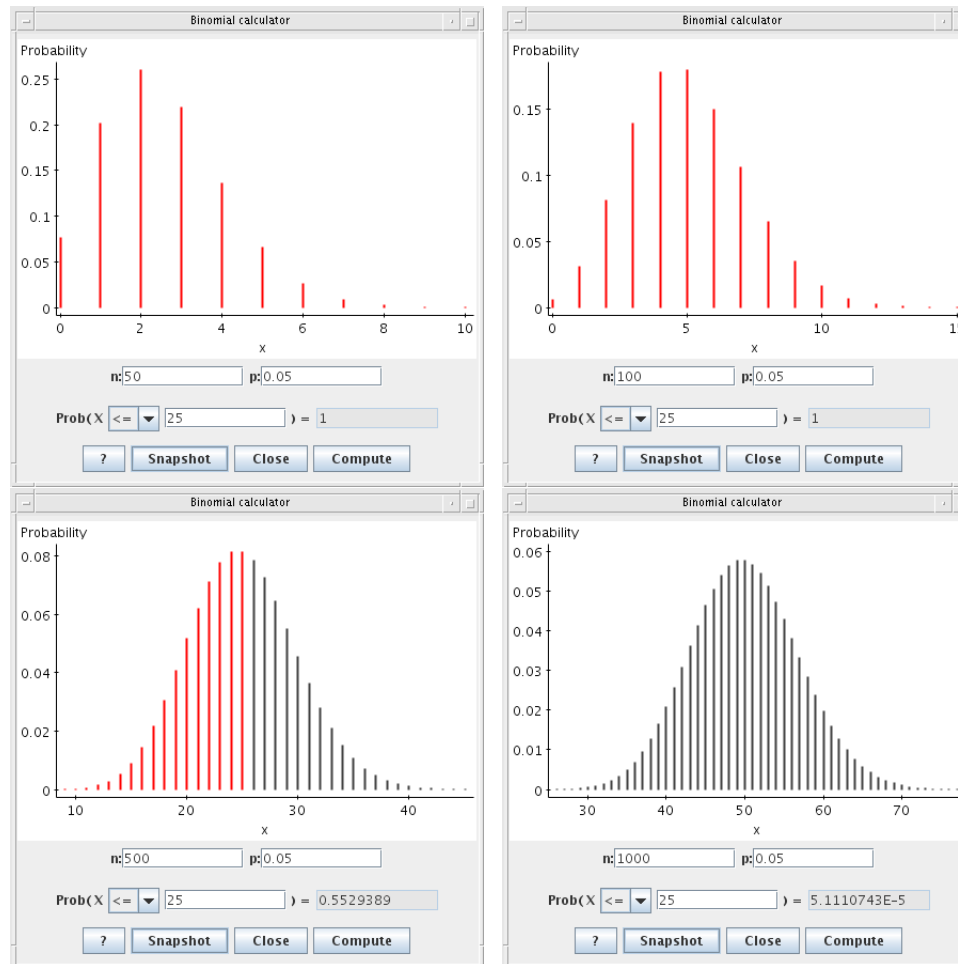
Summary $p = 0.2, n = 4, 50, 200, 1000$



Summary $p = 0.5, n = 4, 25, 50, 100$



Binomial $p = 0.05$ for various n



A comparison of the plots

- When n (number of experiments) is small, some of the binomial distributions are symmetric (if $p = 1/2$) and some are not symmetric and are skewed (if $p \neq 1/2$).
- But as n grows, the center of the distribution gets more and more symmetric (regardless of whether $p = 1/2$ or not).
- However, the closer p is to zero or one the binomial will be highly skewed.
- The closer p is to zero or one, the larger n will have to be to see the “symmetry” and it won’t be symmetric in the tails.
- We illustrate this by plotting the z-transform for different Binomial distributions.

Making a binomial plot $\text{Bin}(n=50, p=0.4)$

- Go to **Cols > New Columns...** Make the first column with the counts (from 0 to 50).

Add columns to 'untitled 46'

Column Name: x

Lock

Data Type: Numeric

Modeling Type: Continuous

Format: Best Width: 12

Use thousands separator (,)

Initialize Data: Sequence Data Number of runs: 1

From: 0

To: 50

Step: 1

Repeat each value N times: 1

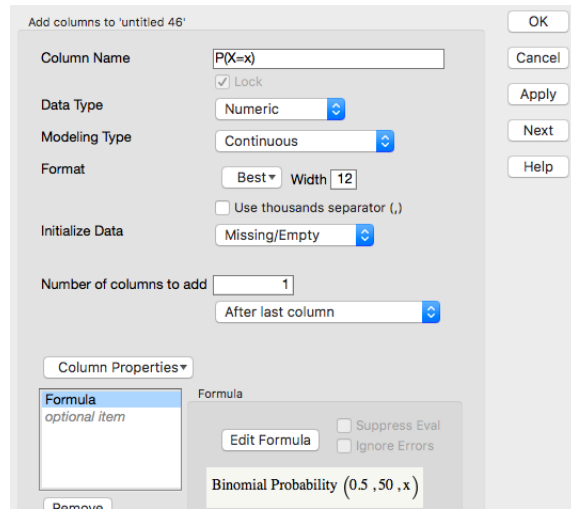
Number of columns to add: 1

After last column

OK Cancel Apply Next Help

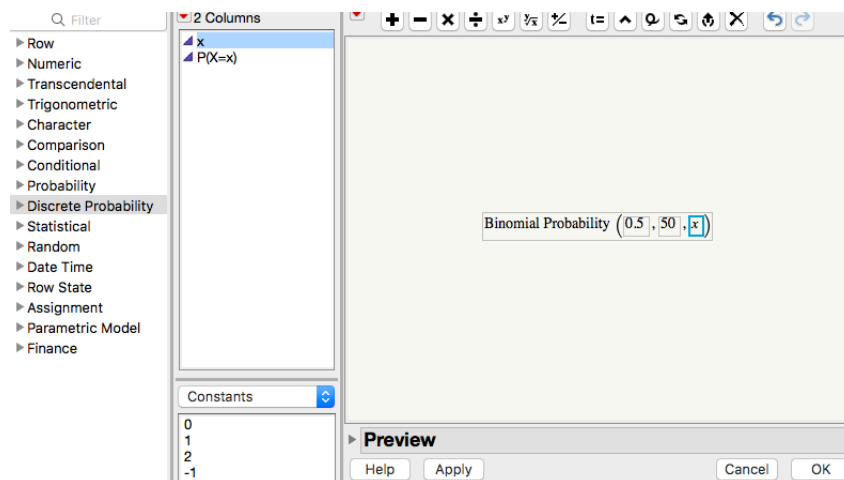
- To make the second column with the probabilities for each count. Go to **Cols > New Columns...** again.

Lecture 8 (MWF) Introducing the normal distribution



But ensure in Column properties you choose the formula option. Once you do this another window will pop up. This is given in the screen shot on the next page.

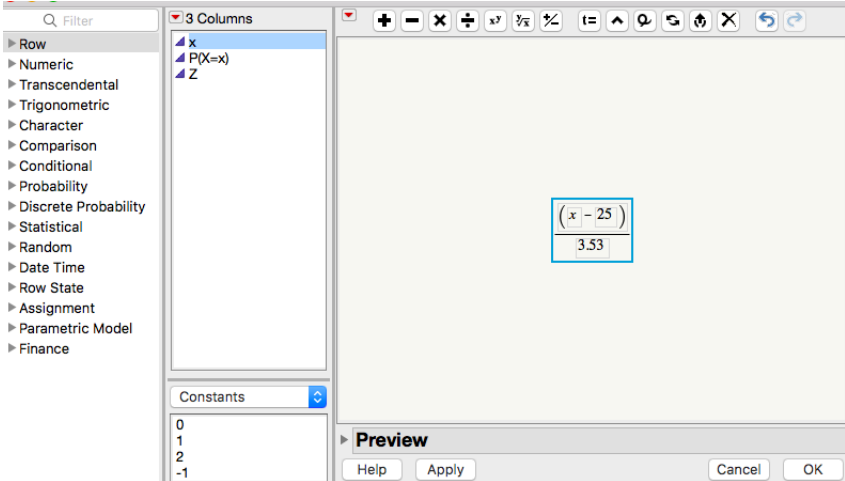
- Choose Discrete Probability (far left of new box) and the Binomial Probability (not Binomial distribution). In the Binomial probability first place the chance of a success (in this example it is 0.5). Then place the number of trials (in this case it is 50, since $S_{50} \in \{0, 1, \dots, 50\}$) and finally highlight the variable X in the first column and drag it over to the Binomial. This will be all the outcomes it evaluates the binomial distribution.



- We create a third column, where we transform the outcomes using the

z-score

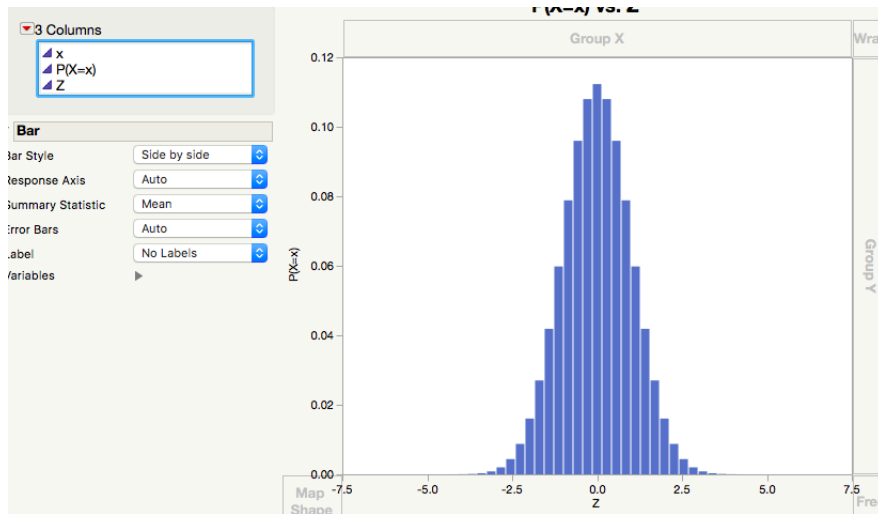
$$z = \frac{\text{outcome} - \text{mean}}{\text{s.d}} = \frac{X - (50/2)}{\sqrt{50/4}}$$



	x	P(X=x)	Z
1	0	8.881784e-16	-7.082152975
2	1	4.440892e-14	-6.798866856
3	2	1.088019e-12	-6.515580737
4	3	1.74083e-11	-6.232294618
5	4	2.045475e-10	-5.949008499
6	5	1.8818369e-9	-5.66572238
7	6	1.4113777e-8	-5.382436261
8	7	8.8715169e-8	-5.099150142
9	8	4.7684403e-7	-4.815864023
10	9	2.2252721e-6	-4.532577904
11	10	9.1236158e-6	-4.249291785
12	11	0.0000331768	-3.966005666
13	12	0.0001078246	-3.682719547
14	13	0.0003151795	-3.399433428
15	14	0.0008329743	-3.116147309
16	15	0.0019991383	-2.83286119
17	16	0.0043731149	-2.549575071
18	17	0.0087462299	-2.266288952

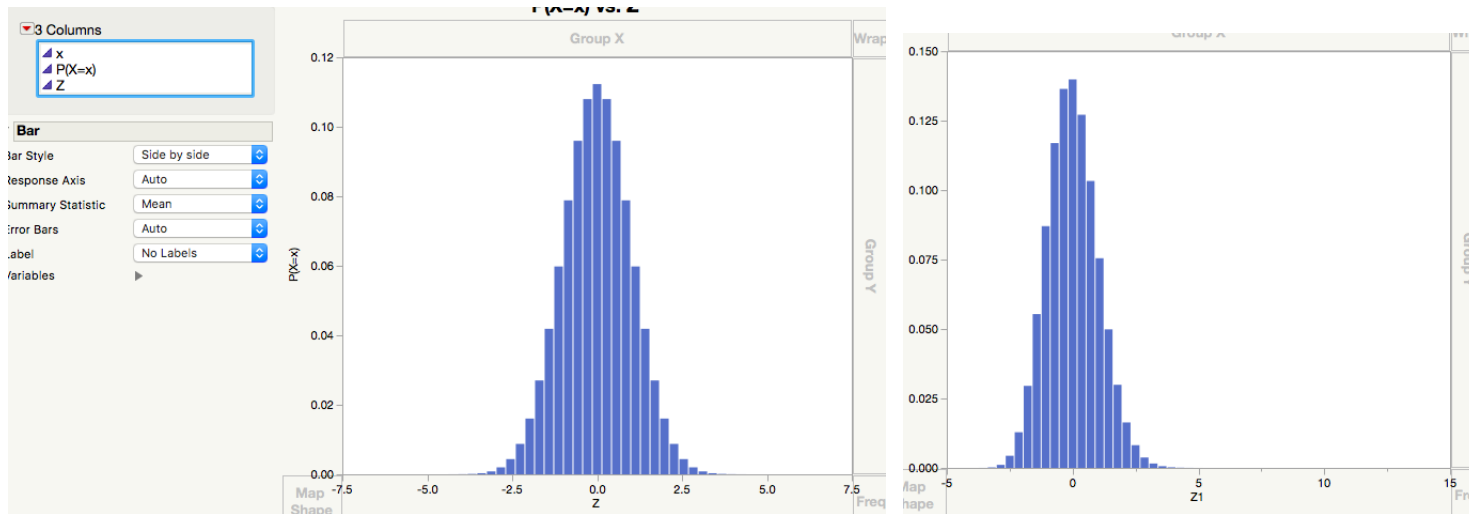
- Finally we plot the z -transformed data against their binomial probabilities

The z-transform of $\text{Bin}(50, p = 0.5)$



- We observe the distribution is symmetric with a certain “bell shapedness”.
- The interesting fact is that as long as n is chosen sufficiently large no matter what p we start with we will always end up with a similar curve.

The z-transform of $\text{Bin}(50, p = 0.5)$ and $\text{Bin}(50, p = 0.2)$



- The above are the z-transforms of the $\text{Bin}(50, 0.5)$ (left) and $\text{Bin}(50, 0.2)$ (right).
- There is still a clear skew in $\text{Bin}(50, 0.2)$, but it is “close” to symmetric.

The plots and the central limit theorem

- The original data is binary (this not even a continuous distribution). The outcomes are either zeros or ones. In terms of notation we observe

$$X_i \in \{0, 1\}.$$

- But the sums of the binary data

$$S_n = \sum_{i=1}^n X_i$$

this is clearly taking on a unique shape as n gets large.

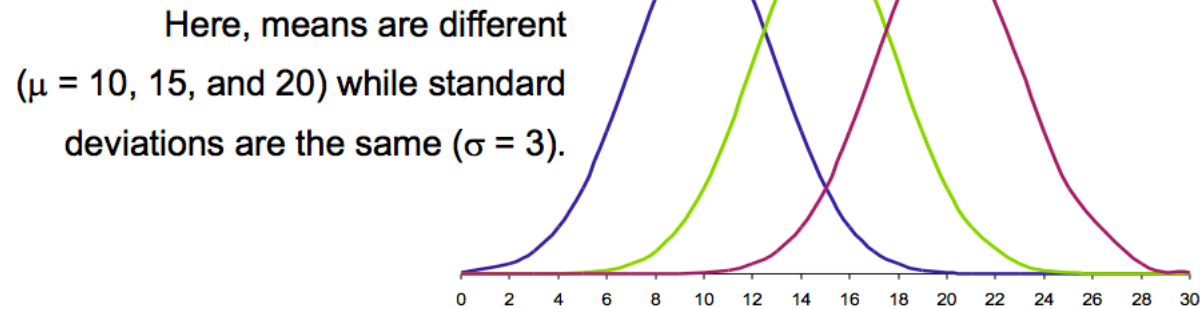
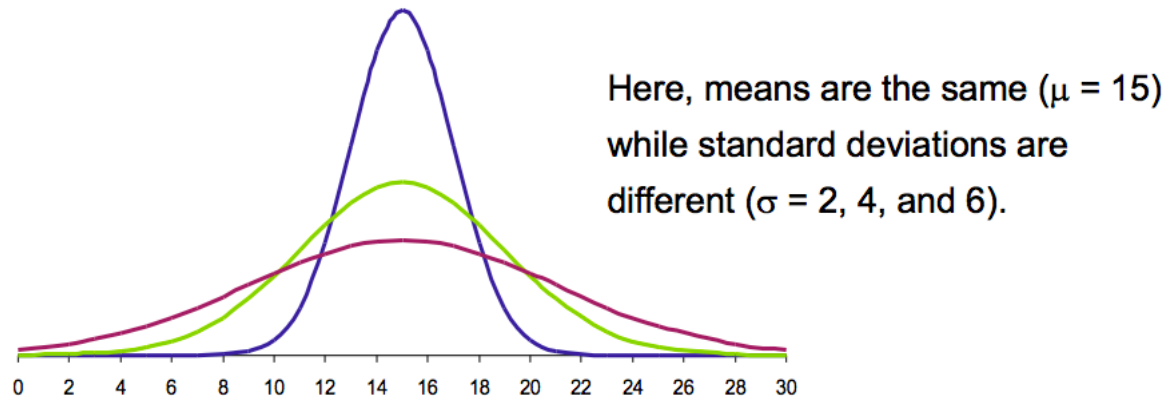
- It is the action of summing or averaging that gives the resulting variable this unique shape.

- The unique shape that we see is the normal distribution.

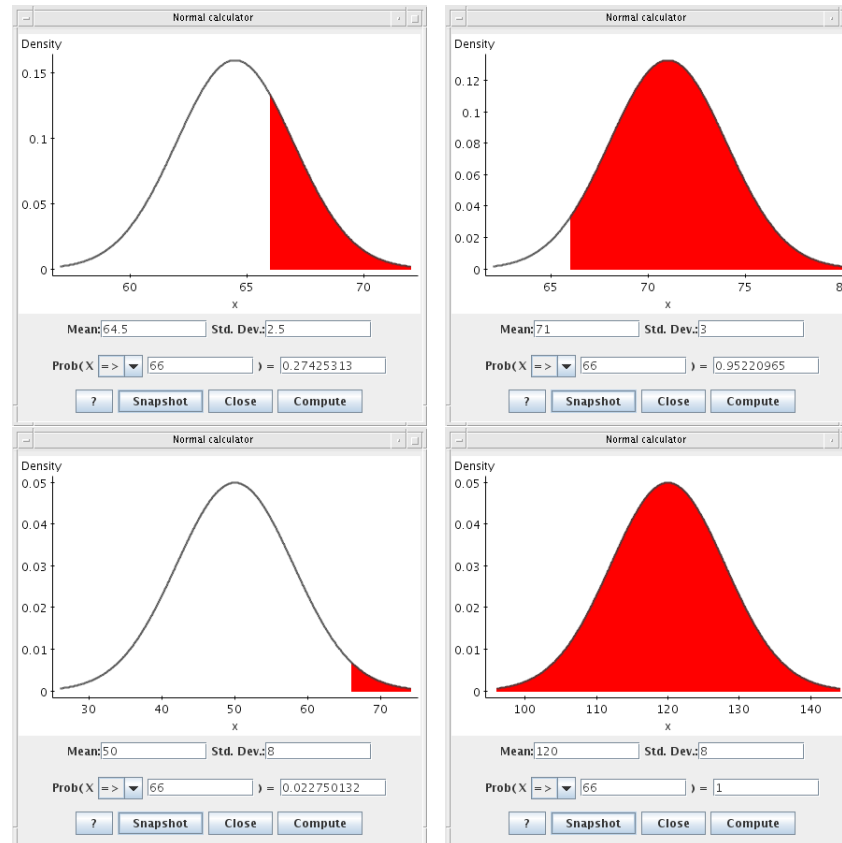
The normal distribution

- It can describe quite well the distribution of random variables of some random variables (but not many).
- Its main application is that it has the elegant property that it describes well the distribution of sample means and sums (we have just seen this in the binomial example)
- The normal distribution is a family of densities which are different but have certain characteristics in common.
- It is completely characterised by two parameters, the mean, μ and standard deviation σ .

The mean and standard deviation



The distribution of heights



Fit the distribution to female human, male human, chimp and giraffe.

z-scores and probabilities

- Suppose the mean weight of healthy calves is 142 pounds and standard deviation is 17 pounds.
- You are presented with a 95 pound baby calf.
- Using the z-score/z-transform (see lecture 4)

$$z = \frac{95 - 142}{17} = -2.76$$

the calf is -2.76 standard deviations from the mean.

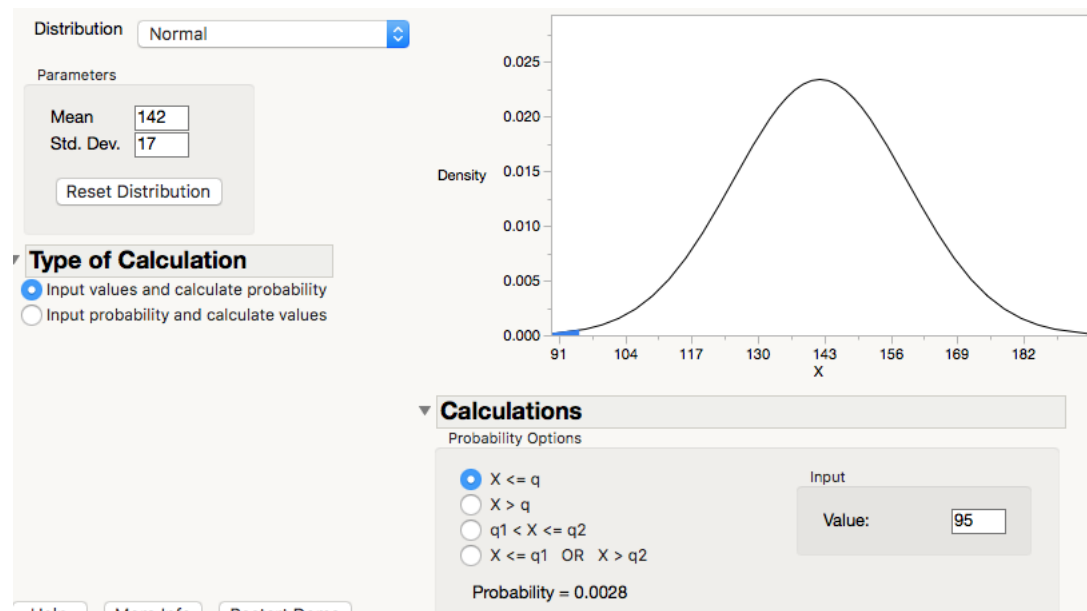
- -2.76 is several standard deviations and we know that the majority of calves lie within one or two standard deviations of the mean.

- Thus a 95 pound calf is quite a rare occurrence amongst healthy calves.
- To find out how rare we need to calculate its percentile.
- A percentile calculation requires a distribution. The weights of baby animals tend to be normal. Thus we assume that the weight of a healthy calf is **normally distributed** with mean 142 and standard deviation 17.
- Using the z-transform we can calculate the probability.

The standard normal - page 1090 of Longnecker and Ott

- The normal tables give the probabilities $P(Z < z)$ in the special case $Z \sim N(0, 1)$ (the so called standard normal):
 - mean is zero ($\mu = 0$)
 - variance is one $\sigma^2 = 1$ (and standard deviation is $\sigma = 1$).
- Look at the normal tables. Suppose we want to use it to evaluate the $P(Z < b)$. The two sides of the table give together b , the inside of the table yields the probability $P(Z < b)$.
- Suppose we want to evaluate $P(Z \leq -2.76)$, since $-2.76 = -2.7 - 0.06$, the first column gives the -2.7 values and first row gives the -0.06 value. We find the -2.7 and -0.06 values and locate the value in the inside of the table where this column and row intersect.

- This intersection point is the probability, that is $P(Z \leq -2.76) = 0.0029 = 0.29\%$.



- The area under the graph is the probability, which corresponds to the value given in the table.

The little calf

- A calf of 95 pounds is

$$z = \frac{95 - 142}{17} = -2.76$$

is -2.76 standard deviation less than the mean of healthy calves. This immediately suggests that the weight is unusual amongst healthy calves.

- If we can assume normality of calf weights, then amongst healthy calves the calf is in the 0.29% percentile. This means she is very light for a healthy calve.
- There are two possible explanations for her height. She is simply an “unusual” healthy calf, or her low weight may indicate an underlying medical issue.

Examples - standard normal

If the random variable is normal, then the z-transform/z-score transforms the normal distribution to one particular normal distribution which has mean zero and variance 1 $N(0, 1)$. This is often called a standard normal. For $Z \sim N(0, 1)$ evaluate (always plot the distribution)

- (a) Evaluate $P(0.6 < Z \leq 1.3)$.
- (b) (i) $P(Z \leq -1.1)$, (ii) $P(Z \leq 0.6)$, (iii) $P(Z \leq 3.0)$, (iv) $P(Z \leq -2.12)$.
- (c) How to interpret $P(Z \leq -1.1)$ and $P(Z \leq 3.0)$?
- (d) (i) $P(Z > -1.1)$, (ii) $P(Z > 0.6)$, (iii) $P(Z > 3.0)$, (iv) $P(Z > -2.12)$.
- (e) (i) $P(-1.1 < Z \leq 0.6)$, (ii) $P(-2.12 < Z \leq 3.0)$, (iii) $P(-2.12 < Z \leq 0)$

Look at the handout http://www.stat.tamu.edu/~suhasini/teaching651/standard_normal_tables.pdf for the solutions.

Calculating probabilities in JMP

Go to Distribution Calculator (follow the same instructions given in lecture 7). Select the normal distribution. Put in the correct mean and standard deviation.

