Objectives

- Density curves
- Measuring center and spread for density curves
- Normal distributions
- The 68-95-99.7 (Empirical) rule
- Standardizing observations
- Calculating probabilities using the standard Normal Table (CIS Chapter 8, p 105 – mainly p114)
- Inverse Normal calculations

Books: OS3: Section 3.1 (entire section). IPS: p242, Section 4.3

Topic: Density Curves

- Learning targets:
  - Understand how to interpret the relative frequency plot of data
  - Understand how to calculate proportions in a sample based on a relative frequency graph.
  - Understand what a density curve is.
  - Understand that the area under a curve corresponds to proportions in a population.
  - Understand that the shapes of density curves can be very different, with many different features.
  - Be able to make a rough sketch of the density curve of a variable and be able to place the mean and standard deviation on it.

Formal: Probabilities

- Anything which varies within a set of variables due to chance is called a random variable. The random variable is often denoted as X. The notation Ω is called the "sample space" and is the set of all possible outcomes of X.

- A random variable can be discrete i.e. taking any one of the values in the set

$$\Omega = \{a, b, c, d, \ldots\}$$

- Or it can be continuous taking any value in the interval

$$[u_1, u_2] \qquad u_1 \text{ and } u_2 \text{ are numbers.}$$

- We associate a probability to all the possible outcomes.
- If it is discrete, then we write

  $$P(X = a)$$

  to denote the probability of the random variable taking the outcome a.
- If it is continuous, then we write

  $$P(a < X \leq b)$$

  to denote the probability of the random variable taking any number between a and b.
- A probability must lie between zero and one.

For the purpose of this course you can treat probability and proportion as the same.

Tossing a six sided die

- The throw of a six-sided die gives rise to the possible outcomes $\{1, 2, 3, 4, 5, 6\}$. Observe that the numbering does not have any real ordering (unless we choose it to), in which case X is a categorical random variable.

- If the the die is completely fair then

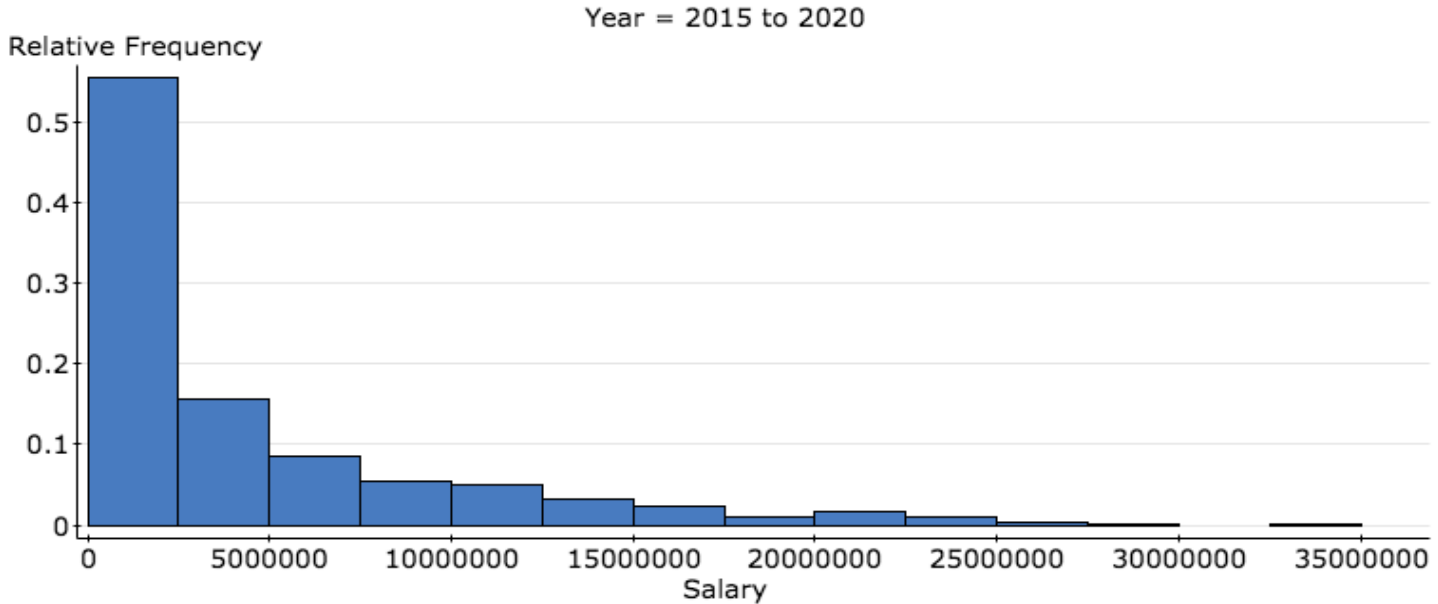
$$P(X = 1) = 1/6, \quad P(X = 2) = 1/6, \quad P(X = 3) = 1/6,$$
$$P(X = 4) = 1/6, \quad P(X = 5) = 1/6, \quad P(X = 6) = 1/6.$$

Reminder: Notation used

Measure	Population (unknown) Parameters (unknown)	Sample Estimates
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s

Histogram and density curves

- If the data is a continuous numerical variable, it can take any value in a given range of numbers
  - Heights of people can lie anywhere between 0.5 meters to 2 meters.
  - Weights of pigs can lie anywhere between 150 pounds to 600 pounds.
- For any **sample** from a continuous numerical variable we plot the *relative frequency histogram*.

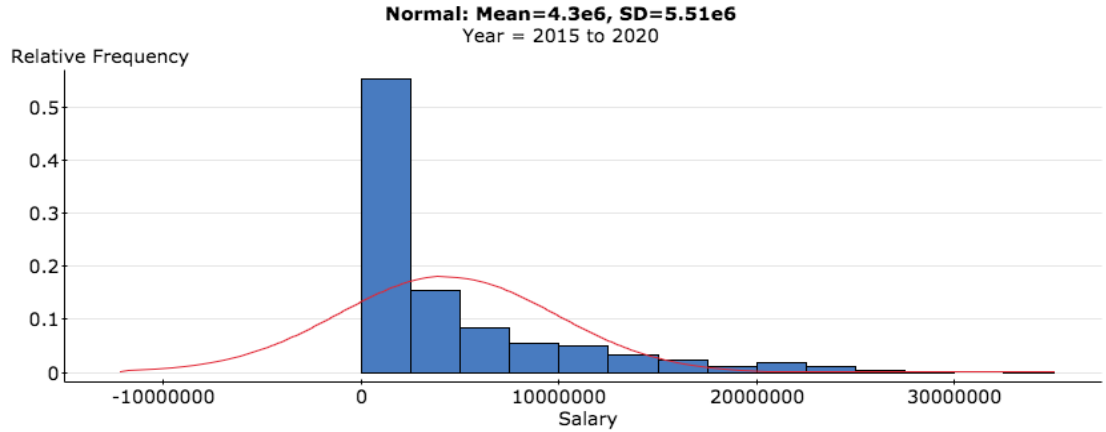


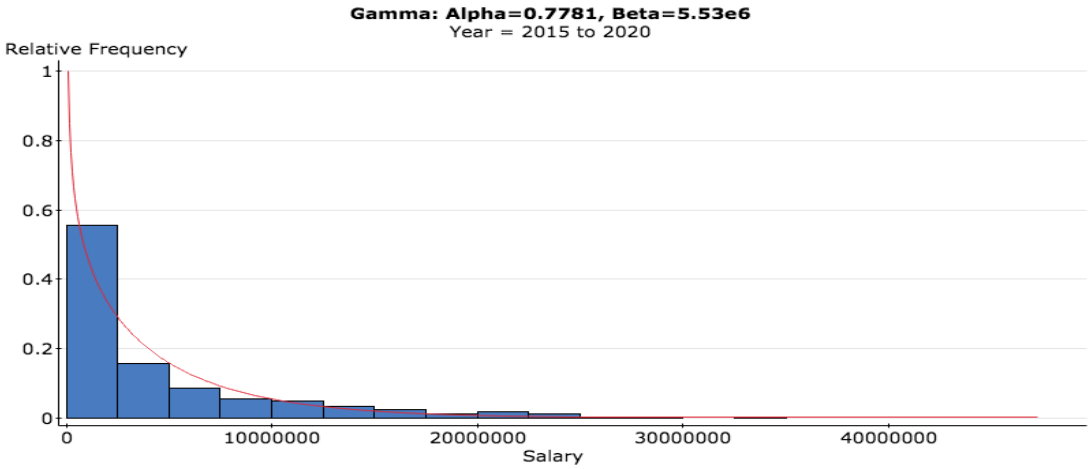
Histogram and density curves

- The “histogram” of the **population** of a continuous random variable is called the *density curve*. It differs from the histogram of a sample in a few important ways:
  - The **height** of the **relative frequency** histogram gives the proportion/chance over any given interval.
  - For sum of all heights is one.
  - The **area** under the **density curve** gives the proportion/chance over any given interval.
  - The area under the entire curve is one.
- As we rarely observe the entire population, we do not know the true density curve. Based on the sample we can often obtain a good estimate using statistical software.

Example: Fitting a density curve

- Load data into Statcrunch.
- Plot a histogram based on your data set.
- In Display options on the histogram menu, there is an option called Overlay distrib.
- In this menu there is a list of density shapes with different shapes. You can overlay your histogram to see which best fits your data.
- We give some examples on the next slide.





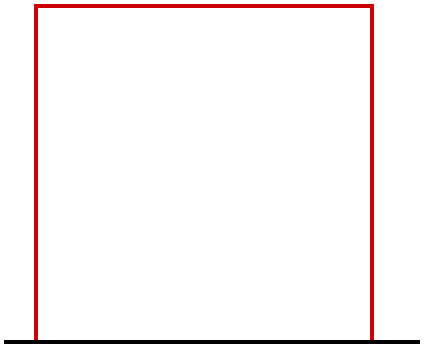
We plot two different densities over the relative frequency of base ball salaries. Which fits the data better?

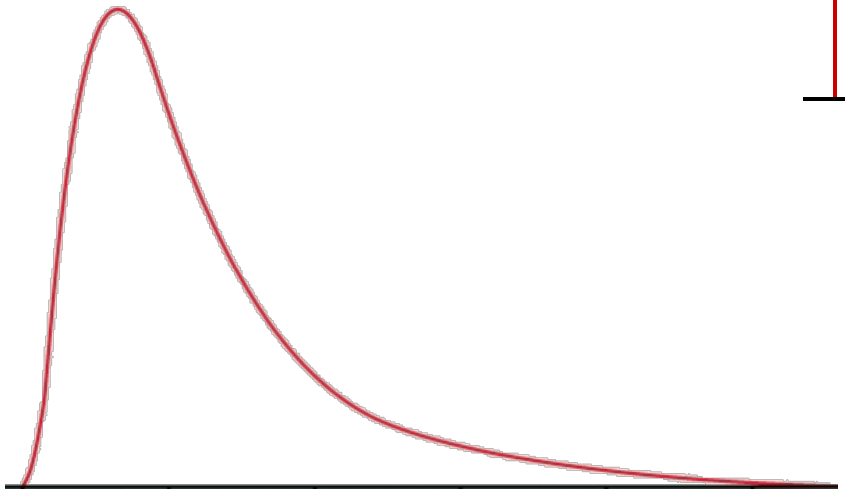
Plotting densities and visualization

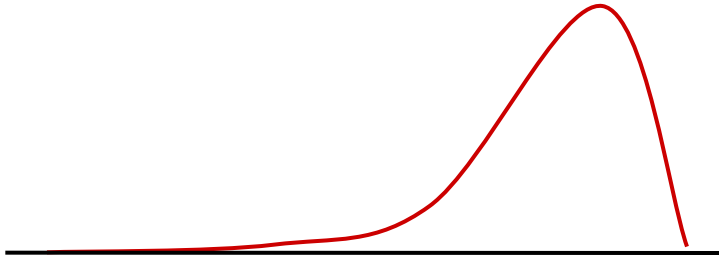
- Make a sketch of what you think the density curve of human heights are (use mean 67 inches and standard deviation 7 inches as an aid).
- What is the area below the **entire** curve?
- On the plot show the proportion of human heights less than 60 inches.
- On the plot show the proportion of human heights greater than 75 inches.
- On the plot show the proportion of human heights lying between 60 and 75 inches.

The density of heights

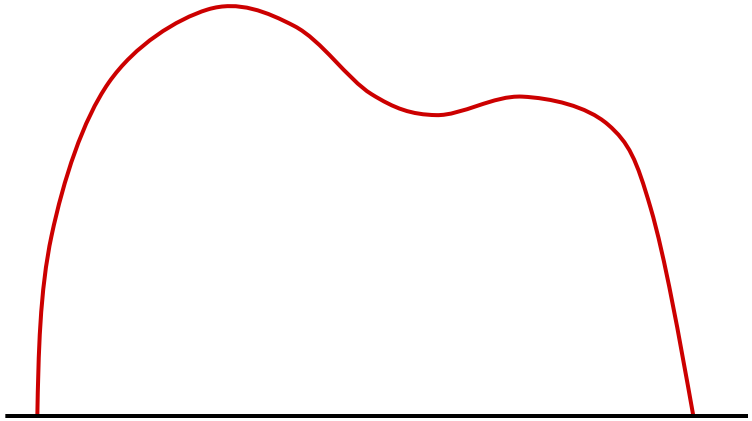
Density curves come in any imaginable shape.







The only restrictions is that the y-axis **cannot take negative values** and the **area below the curve is one.**



Topic: The normal distribution

- Learning targets:
  - Understand that the normal distribution is one particular family of density curves.
  - Understand that the normal distribution is determined by its mean and standard deviation.
  - Understand the main features of the normal distribution.
  - Calculate z-scores and percentiles using the normal distribution (using both tables and statcrunch). Understand that all these calculations are based on the **assumption** the data is normal. If the data is not normal then the calculated percentile can be wrong.
  - Make comparisons using percentiles.
- OS3, page 128 onwards.
- Free normal calculators http://onlinestatbook.com/2/calculators/normal_dist.html

The normal family of density plots

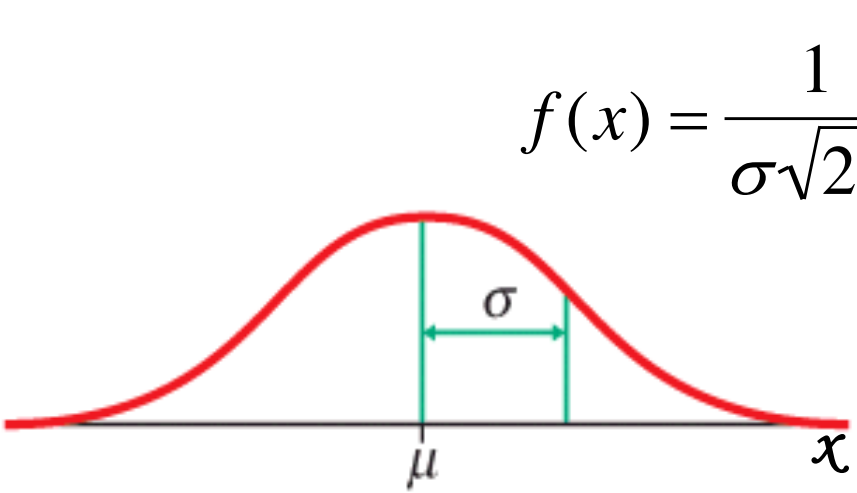
- We now introduce a family of density functions which are extremely useful in statistics. It is called the **normal distribution**.
- Here are some reasons that they are important in statistics
  - **Some** variables (but **not** all) have a density which is close to a normal distribution. These include biological measurements, some type of exam scores etc.
  - Calculations using the normal distribution are extremely simple.
  - The central limit theorem states that the distribution of average (sample mean) is “close” to normal. This result is used to construct confidence intervals and tests.

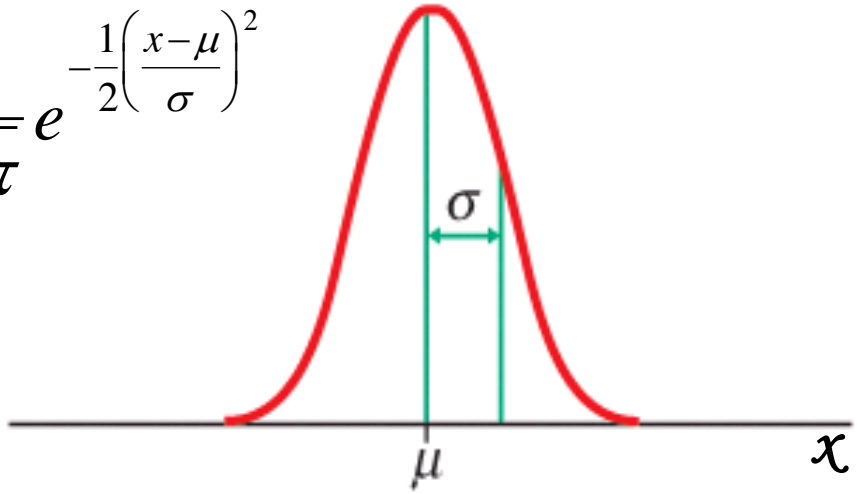
Definition: Normal density

Normal distributions are a *family* of **symmetrical**, "bell-shaped" density curves defined by a mean μ (*mu*) and a standard deviation σ (*sigma*). We denote a normal distribution by Normal(μ,σ) or N(μ,σ).

The formula for the density curve is somewhat complicated:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

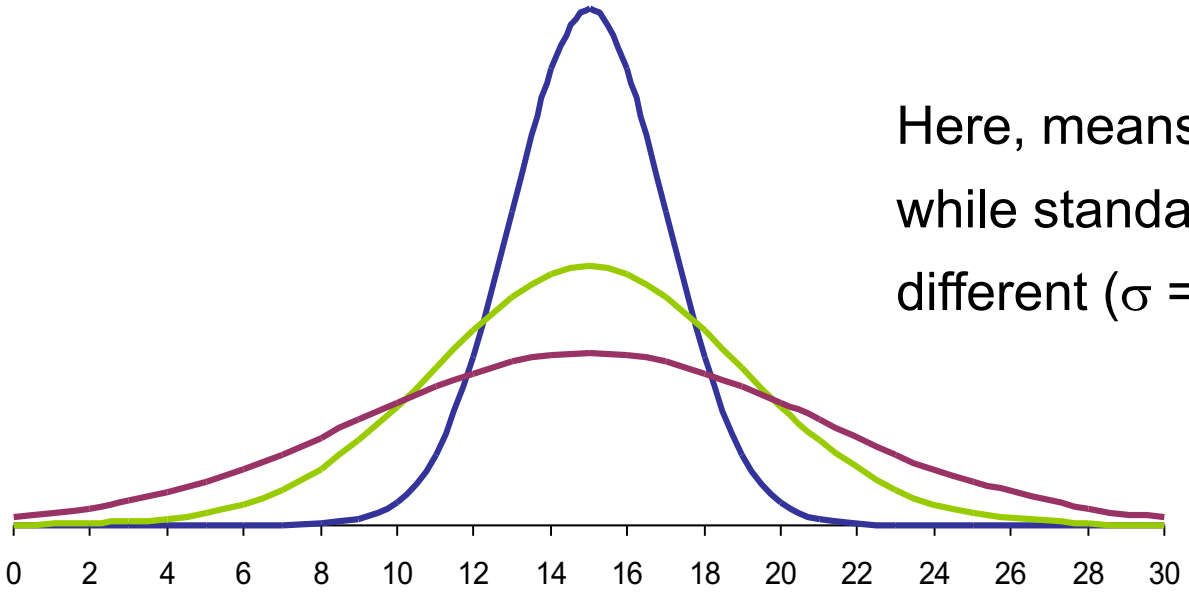




e = 2.71828… the base of the natural logarithm

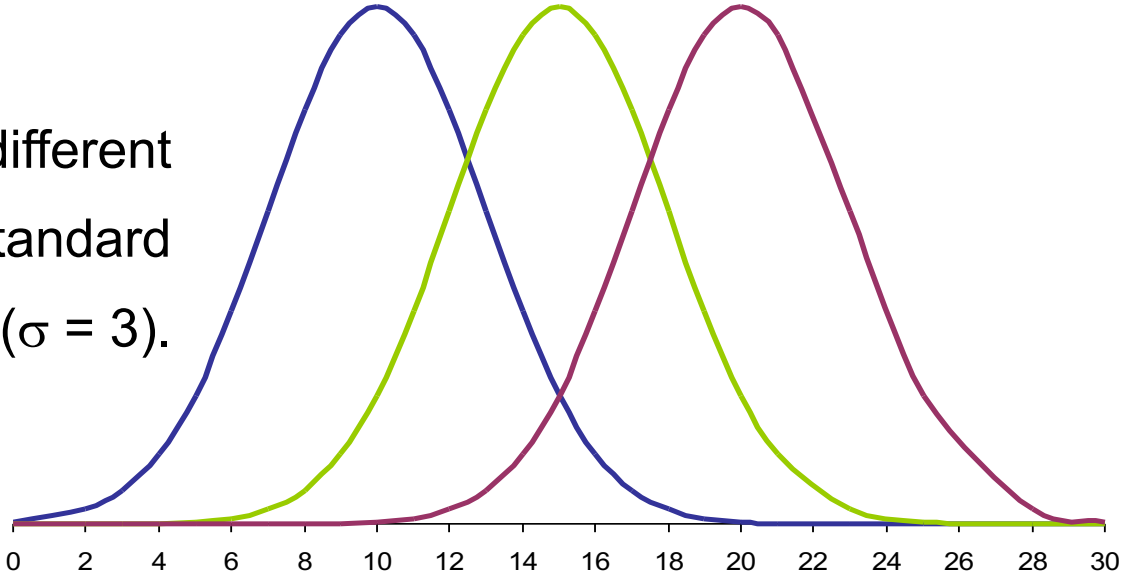
π = pi = 3.14159…

Examples of normal density curves



Here, means are the same (μ = 15) while standard deviations are different (σ = 2, 4, and 6).

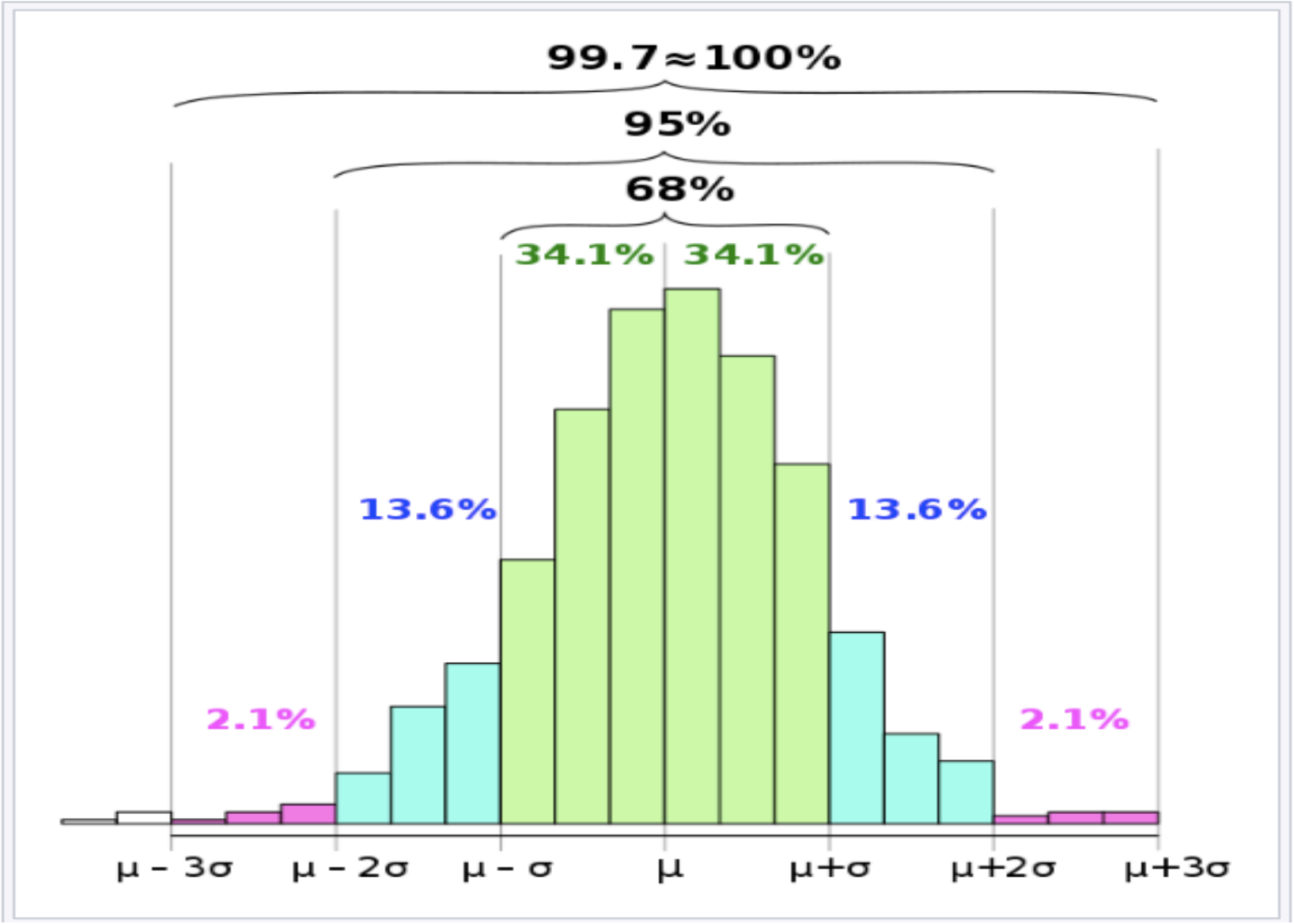
Here, means are different (μ = 10, 15, and 20) while standard deviations are the same (σ = 3).



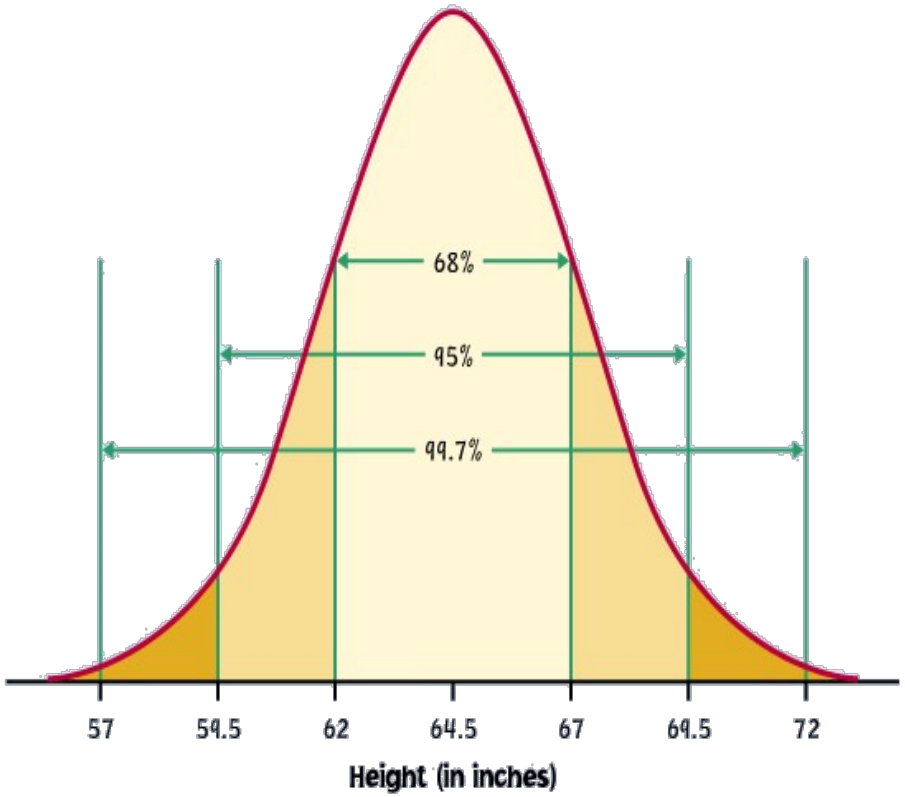
Statcrunch: Overlay normal density

- Load the calf data into Statcrunch. Here the aim is to compare the histogram of calf weights to the normal density curve.
  - Graphics -> Histogram
    - Select a weight variable (such as 8 weeks)
    - Select relative frequency (in Type options).
    - In Overlay distrib choose normal density. Statcrunch will calculate the sample mean and standard deviation of the weights and use this to center the normal density curve.
- Remember the fit won't be perfect (even when the true population density is normal).

Empirical rule: A crude rule for checking Normal Distributions



- Using the symmetry of the normal distribution and the plot on the right as a guide we make the following observations:
- About 34% of all observations are within **1** standard deviation to the **left** of the mean.
- About 47.5% of all observations are within **2** standard deviations to the **left** of the mean.



mean μ = 64.5 standard deviation σ = 2.5

Normal(μ, σ) = Normal(64.5, 2.5)

Important observation

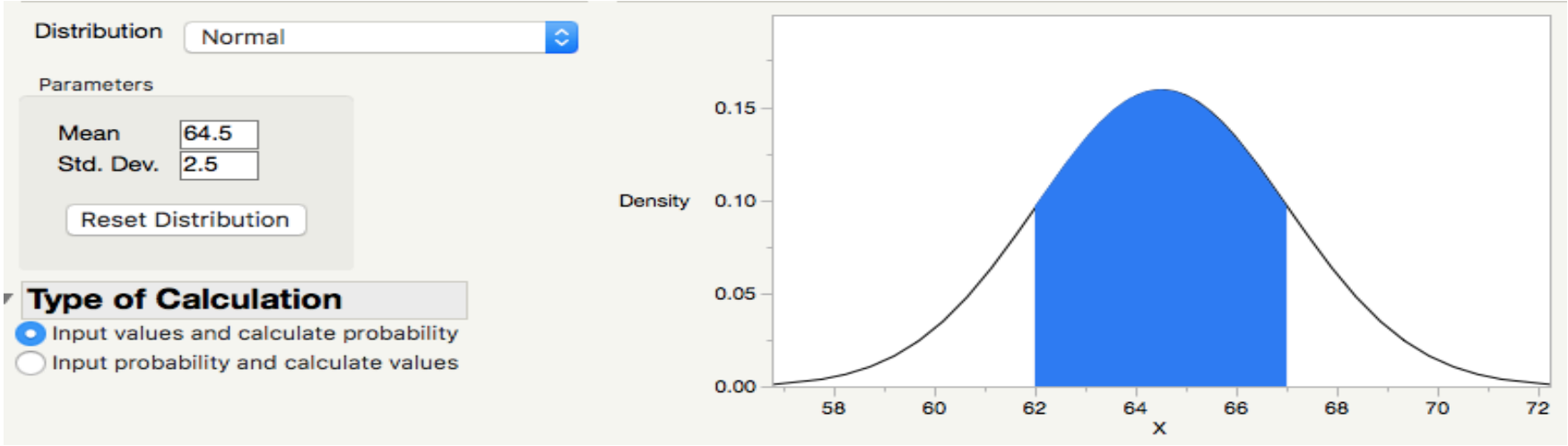
- For a normal distribution: 68% of the observations are within **one standard deviation** of the mean.
- **This does <u>not</u> mean**
  - $2 \times 68 = 136\%$ of the observations lie within **two standard deviations** of the mean. In fact, **95%** of the observations are within **two standard deviations** of the mean.
  - $0.5 \times 68 = 34\%$ of the observations lie within **half a standard deviation** of the mean. In fact, **38%** of observations lie within half a standard deviation of the mean.

Z-transform

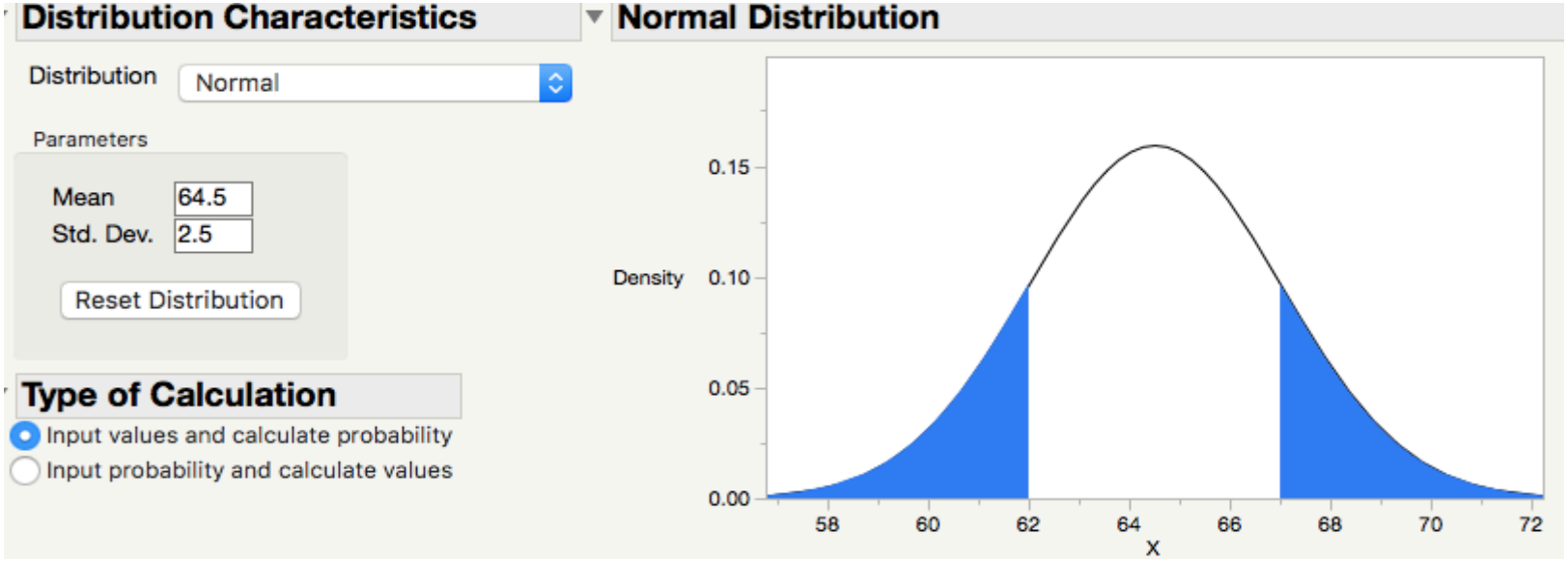
- The z-transform measures of how many standard deviations the observation is from the mean. We explain why below:

$$\text{z-transform} = \frac{\text{data - mean}}{\text{st.dev}}$$

- All observations that are within one standard deviation of the mean have a z-score less than one (blue region below).



- Conversely, all observation more than one standard deviation from the mean have a z-score greater than one (blue plot below).



Do baseball salaries satisfy the check for normality?

Summary statistics:

Column	n	Mean	Variance	Std. dev.
SalaryMillion	817	4.3012761	30.317996	5.506178

Applying the empirical rule to determine healthy calf.

- You are presented with an 8 week calf whose weight is 95 pounds. Is he healthy?
- Calculating the z-transform:

$$\sigma = 17, \quad \mu = 142.6, \quad z = \frac{95 - 142.6}{17} = -2.8$$

- The calf is -2.8 standard deviations **below** the mean.
- -2.8 is quite far from where most weights lie.
- Suppose the weights are normally distributed:

Negative and positive z-transforms

- If the data is **less** than the mean, then the z-transform will be negative.
  - Example: The 95 pound calf, in the previous example, was 2.8 standard deviations less than (to the **left** of) the mean of 142 pounds. Its z-transform is **-2.8**.
- If the data is greater than the mean, then the z-transform will be positive.
  - Example: A 190 pound calve is 2.8 standard deviations to the **right** of the mean (142 pounds).

    Its z-transform is **2.8**.
- In both examples the calves are 2.8 standard deviations from the mean, but in completely different directions.

Z-scores and the normal density

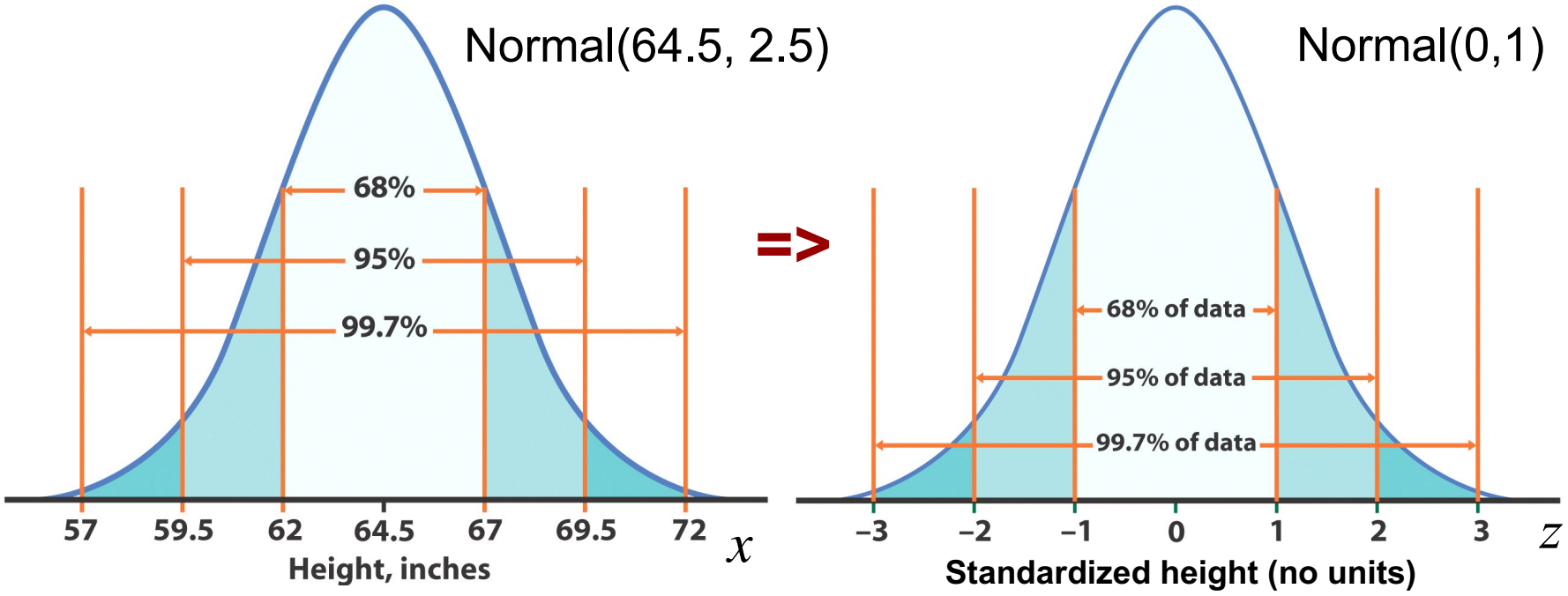
- **Example**: Suppose the heights of women are close to normally distributed with mean 64.5 inches and standard deviation 2.5 inches.
- **Question:** A women has a height of 71 inches, is she exceptionally tall?
- **Answer:** The z-transform calculates how close 71 is to the mean but takes into account the spread of heights:

$$z = \frac{71 - 64.5}{2.5} = 2.6$$

- She is tall, but what is the exact percentile? To do this we need to calculate the area to the LEFT of 71 on the density curve.

The standard Normal distribution

The z-transform transforms Normal(μ,σ) curve into the **standard normal** curve: Normal(0,1).

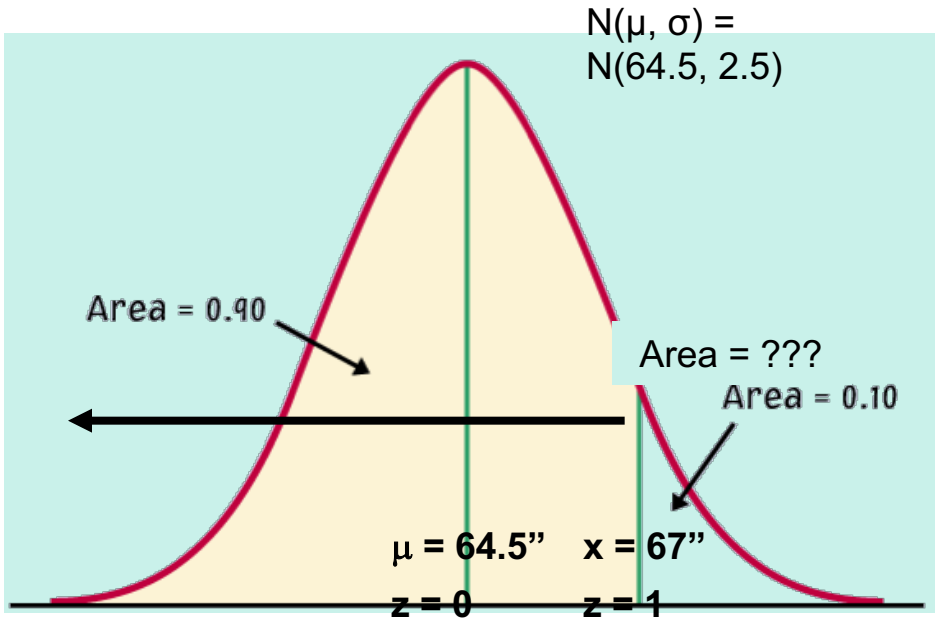


For each *x* we calculate a new value, ***z*-score**.

Calculation: Women heights

Women's heights follow the N(64.5",2.5") distribution. What percent of women are *shorter* than 71 inches tall?

mean μ = 64.5"
standard deviation σ = 2.5"
x (height) = 71"



Always draw a picture of your problem!

We calculate *z*, the standardized value of *x*:

$$z = \frac{x - \mu}{\sigma}, \quad z = \frac{71 - 64.5}{2.5} = 2.6 \quad 2.6 \text{ s.d. above the mean}$$

To find the percent of women are shorter than 71 inches tall, we need to find the *area to the left* of *z* = 2.6. For this, we must use a special table.

Standard Normal p	
z	.00
0.0	.5000
0.1	.5398
0.2	.5793
0.3	.6179
0.4	.6554
0.5	.6915
0.6	.7257
0.7	.7580
0.8	.7881
0.9	.8159
1.0	.8413
1.1	.8643
1.2	.8849
1.3	.9032
1.4	.9192
1.5	.9332
1.6	.9452
1.7	.9554
1.8	.9641
1.9	.9713
2.0	.9772
2.1	.9821
2.2	.9861
2.3	.9893
2.4	.9918
2.5	.9938
2.6	.9953

To find the percentile. Go to the normal tables at https://www.stat.tamu.edu/~suhasini/teaching301/zTable.pdf

Look for 2.6 on the row and 0.00 (2.6+0.00 = 2.6). Search for the intersection. It gives you 0.9953.

Conclusion:

99.53% of women are shorter than 71".

By subtraction, 1 – 0.9953, or 0.46% of women are taller than 71".

She is in the **99.53 percentile** or in the **top 0.47 percentile**.

Always make the plot

Using the standard Normal table

Table A gives the area under the standard Normal curve to the *left* of any z value.

TABLE A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.022	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0	.0233
−1.8		.0351	.0344	.0336	.0329	.0322	.0314	.0307	.030	.0294
−1.7		.0436	.0427	.0418	.0409	.0401	.0392	.0384		
		.537	.0526	.0516	.0505	.0495	.0485	.0475		
		.655	.0643	.0630	.0618	.0606	.0594	.0582		
		.793	.0778	.0764	.0749	.0735	.0721	.0708		

(…)

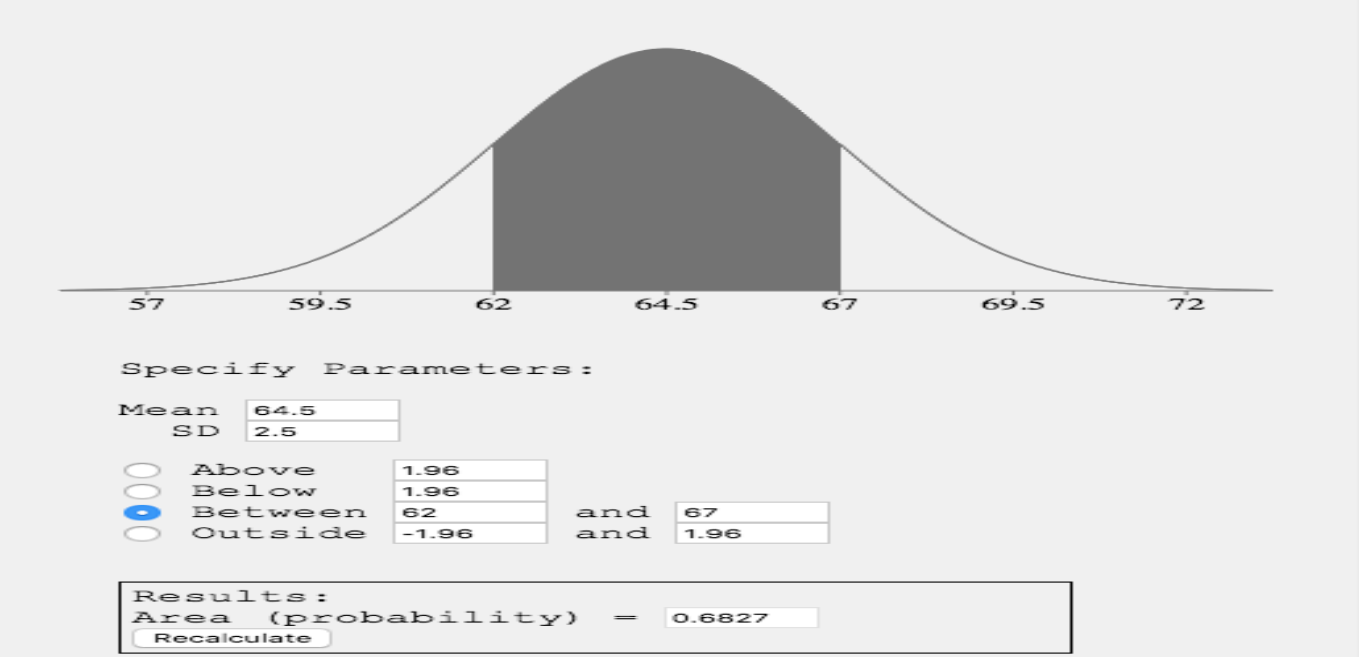
.0082 is the area under N(0,1) left of z = -2.40

.0080 is the area under N(0,1) left of z = -2.41

0.0069 is the area under N(0,1) left of z = -2.46

Probability calculators

- We can calculate probabilities using Statcrunch.
  - Stat -> Calculators -> Select the normal distribution.
  - Here you can choose the mean and standard deviation and calculate the area on the left or right. This area corresponds to the proportion of the population less than or greater than a value.
- http://onlinestatbook.com/2/calculators/normal_dist.html



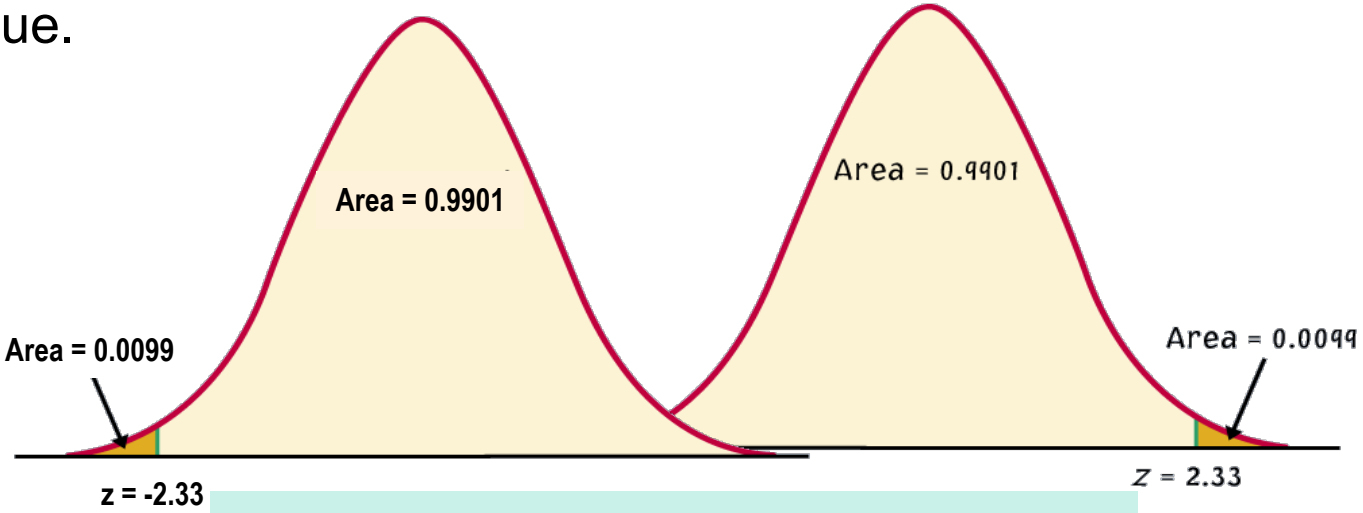
Question Time

- Assume that female heights are normally distributed with a mean height of 64.5 inches and standard deviation is 2.5 inches. What proportion of females have a height **less** than 63 inches (hint: use the first page of the tables and not the second)?

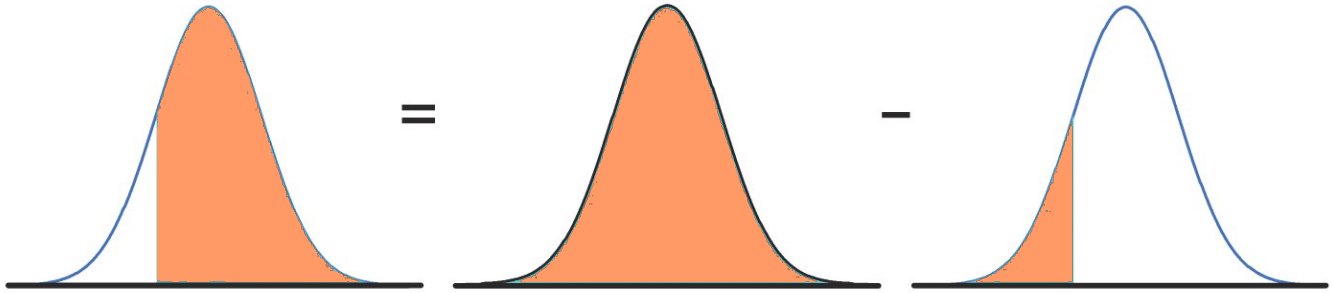
- A. 60%
- B. 76%
- C. 27%
- D. -60%

Tips on using Table A

Because the Normal distribution is symmetric, there are 2 ways that you can calculate the area under the standard Normal curve to the right of a z value.



area to right of z = area to left of $-z$



area to right of z = 1 – area to left of z

Symmetry of the normal

- Compare the first and second page of the z-tables

z	.00	.01	.02
−3.4	.0003	.0003	.0003
−3.3	.0005	.0005	.0005
−3.2	.0007	.0007	.0006
−3.1	.0010	.0009	.0009
−3.0	.0013	.0013	.0013
−2.9	.0019	.0018	.0018
−2.8	.0026	.0025	.0024
−2.7	.0035	.0034	.0033
−2.6	.0047	.0045	.0044
−2.5	.0062	.0060	.0059
−2.4	.0082	.0080	.0078
−2.3	.0107	.0104	.0102
−2.2	.0139	.0136	.0132
−2.1	.0179	.0174	.0170
−2.0	.0228	.0222	.0217
−1.9	.0287	.0281	.0274
−1.8	.0359	.0351	.0344
−1.7	.0446	.0436	.0427
−1.6	.0548	.0537	.0526
−1.5	.0668	.0655	.0643
−1.4	.0808	.0793	.0778
−1.3	.0968	.0951	.0934
−1.2	.1151	.1131	.1112
−1.1	.1357	.1335	.1314
−1.0	.1587	.1562	.1539
−0.9	.1841	.1814	.1788
−0.8	.2119	.2090	.2061
−0.7	.2420	.2389	.2358
−0.6	.2743	.2709	.2676
−0.5	.3085	.3050	.3015
−0.4	.3446	.3409	.3372
−0.3	.3821	.3783	.3745
−0.2	.4207	.4168	.4129
−0.1	.4602	.4562	.4522
−0.0	.5000	.4960	.4920

z	.00	.01	.02
0.0	.5000	.5040	.5080
0.1	.5398	.5438	.5478
0.2	.5793	.5832	.5871
0.3	.6179	.6217	.6255
0.4	.6554	.6591	.6628
0.5	.6915	.6950	.6985
0.6	.7257	.7291	.7324
0.7	.7580	.7611	.7642
0.8	.7881	.7910	.7939
0.9	.8159	.8186	.8212
1.0	.8413	.8438	.8461
1.1	.8643	.8665	.8686
1.2	.8849	.8869	.8888
1.3	.9032	.9049	.9066
1.4	.9192	.9207	.9222
1.5	.9332	.9345	.9357
1.6	.9452	.9463	.9474
1.7	.9554	.9564	.9573
1.8	.9641	.9649	.9656
1.9	.9713	.9719	.9726
2.0	.9772	.9778	.9783
2.1	.9821	.9826	.9830
2.2	.9861	.9864	.9868
2.3	.9893	.9896	.9898
2.4	.9918	.9920	.9922
2.5	.9938	.9940	.9941
2.6	.9953	.9955	.9956
2.7	.9965	.9966	.9967
2.8	.9974	.9975	.9976
2.9	.9981	.9982	.9982
3.0	.9987	.9987	.9987
3.1	.9990	.9991	.9991
3.2	.9993	.9993	.9994
3.3	.9995	.9995	.9995
3.4	.9997	.9997	.9997

The standard normal is symmetric at zero.
The proportion to the left of zero is 50%.
The proportion to the right of zero is 50%.

Question Time

- On the standard normal z-tables, what is the proportion to the **right** of **z = -1.43**?
  - (A) 99.57%
  - (B) 92.36%
  - (C) 7.64%
  - (D) 1.43%
  - (E) -1.43%

Question Time

- Assume that female heights are normally distributed with a mean height of 64.5 inches and standard deviation is 2.5 inches. What proportion of females have a height more than 62 inches?

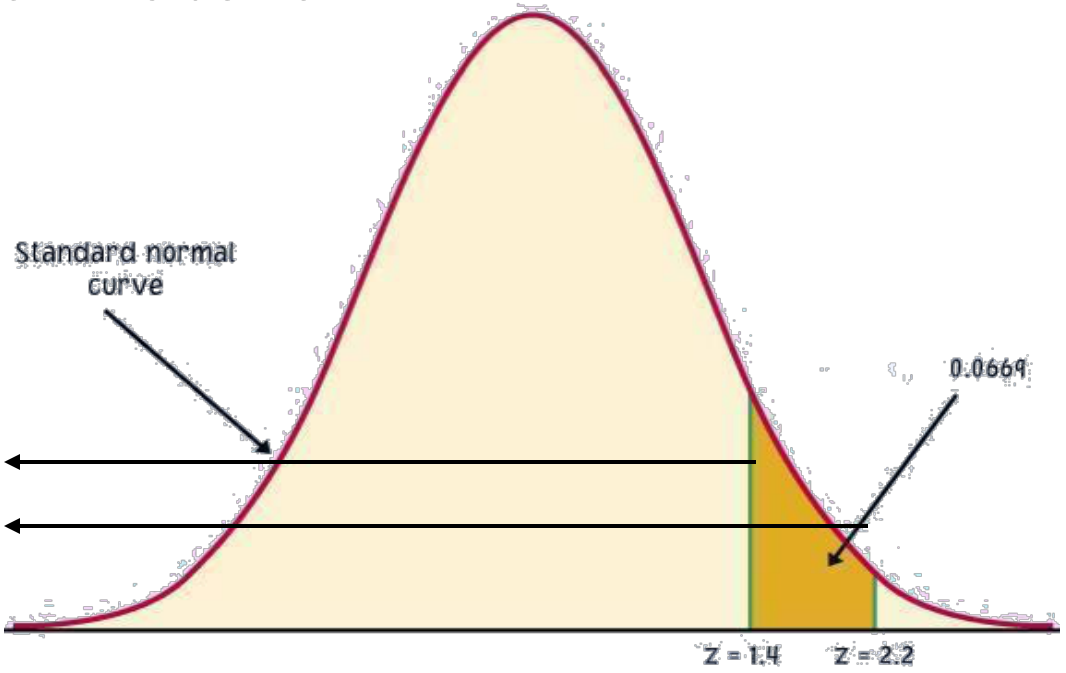
- A. 15.87%
- B. 99%
- C. 69.15%
- D. 1%
- E. 84.13%

Tips on using Table A

To calculate the area between two *z*- values, first get the area under Normal(0,1) to the left of each *z*-value from Table A.

Then subtract the smaller area from the larger area.

A common mistake made by students is to subtract the z values instead of subtracting the areas.



The area between z_1 and z_2 is the area left of z_1 minus the area left of z_2.

Calculation Practice:

- Question: What proportion of females have height **between** 60.2 to 70 inches?

- Answer: Calculate the z-transform corresponding to 60.2

$$z_1 = \frac{60.2 - 64.5}{2.5} = -1.72$$

Standard Normal probabilities

z	.00	.01	.02
−3.4	.0003	.0003	.0003
−3.3	.0005	.0005	.0005
−3.2	.0007	.0007	.0006
−3.1	.0010	.0009	.0009
−3.0	.0013	.0013	.0013
−2.9	.0019	.0018	.0018
−2.8	.0026	.0025	.0024
−2.7	.0035	.0034	.0033
−2.6	.0047	.0045	.0044
−2.5	.0062	.0060	.0059
−2.4	.0082	.0080	.0078
−2.3	.0107	.0104	.0102
−2.2	.0139	.0136	.0132
−2.1	.0179	.0174	.0170
−2.0	.0228	.0222	.0217
−1.9	.0287	.0281	.0274
−1.8	.0359	.0351	.0344
−1.7	.0446	.0436	.0427

Look look the z-tables. Since
-1.72 = -1.7 - 0.02, use the intersection of the column and row inside the table.

We see from the table it is 0.0427.

- Calculate the z-transform corresponding to 70

$$z_1 = \frac{70 - 64.5}{2.5} = 2.2$$

Standard Normal

z	.00
0.0	.5000
0.1	.5398
0.2	.5793
0.3	.6179
0.4	.6554
0.5	.6915
0.6	.7257
0.7	.7580
0.8	.7881
0.9	.8159
1.0	.8413
1.1	.8643
1.2	.8849
1.3	.9032
1.4	.9192
1.5	.9332
1.6	.9452
1.7	.9554
1.8	.9641
1.9	.9713
2.0	.9772
2.1	.9821
2.2	.9861

Look up 2.2 in the table. Since 2.2 = 2.2 + 0.00 we find the intersection of the row and column inside the table.

It is 0.9861

- Since we want the proportion of heights between 60.2 and 70 we take the differences of the areas.
- The answer is 0.9861 – 0.0427 = 0.9434.
- **94.34% of women are between 60.2 and 70 inches.**

- Since we want the proportion of heights between 60.2 and 70 we take the differences of the areas.
- The answer is 0.9861 – 0.0427 = 0.9434.
- **94.34% of women are between 60.2 and 70 inches.**

Question Time

- Assume that male heights are normally distributed with a mean height of 67 inches and standard deviation is 3.5 inches. What proportion of males have a height between 62 and 65.4 inches?

- A. 26.11%
- B. 24.7%
- C. 64%
- D. 1.36%
- E. 2.64%

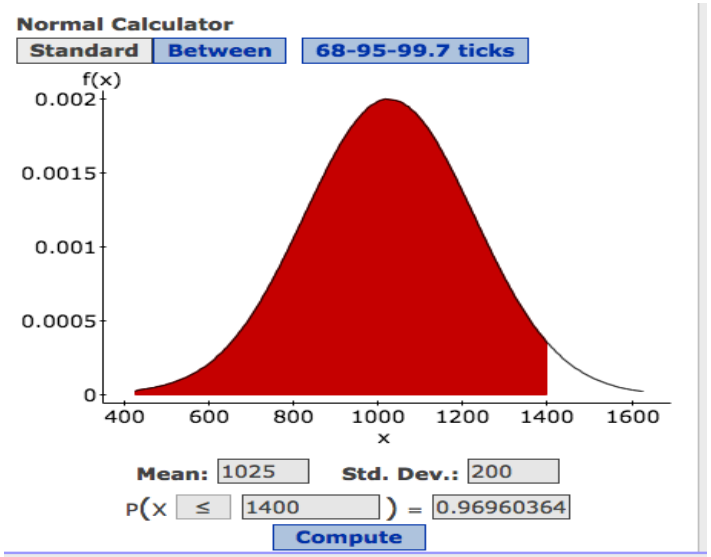
Calculation: Comparing different exams using percentiles

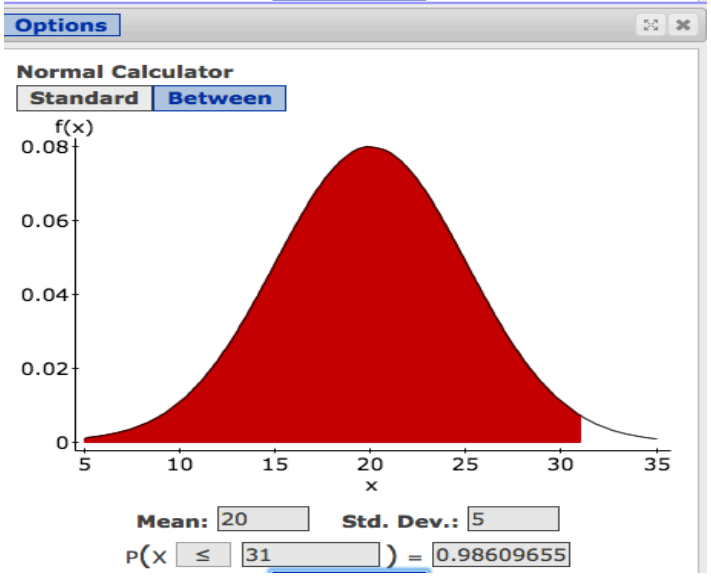
- There are various ways to gain entrance into Texas A&M: SATs and ACTs.
- The ACTs range from 1-36 (very different to SATs).
- How to compare students who have taken different exams?
- The easiest way is by comparing their percentiles, if one student A is in the top 10% SAT scores whereas student B is in the top 5% ACT scores. It is clear that student B did better in their exams.
  - **Important** The comparison is based on the assumption that groups of the same ability took the exams.

- **Question**: SATs have mean score 1025 and standard deviation 200, whereas ACT scores have mean 20 and standard deviation 5.
  - Betty scores 1400 on her SATs,
- Joan scores 31 on her ACT
- Which student did better?
- **Answer** Both students did better than the mean. But to compare the performance of each student we need know how well they did compared with everyone else who took the same exam.

- This requires us to know the scores **of all students** who did both exams. Often such data is unavailable.
- Instead we **assume** that SAT and ACT scores are close to normally distributed. We assume
  - SAT scores are almost normally distributed with mean 1025 and standard deviation 200.
  - ACT scores are close normally distributed with mean 20 and standard deviation 5.
- Assuming the scores are normally distributed allows us to make comparisons between distribution using only the knowledge of the mean and standard deviation of scores.

Making comparisons





Placing both plots in the same scale allows us to make a comparison.

From the picture, it seems that Joan is further to the right.

Joan is in a higher percentile

- The calculation: We first make a z-score for both Betty and Joan:
  - Betty's z-score is z = = 1.875
  - Joan's z-score is z = = 2.2.
  - Using the normal tables
    - Betty is in the 96.7 percentile,
    - Joan is in the 98.6 percentile.
  - Joan did slightly better than Betty, since only 1.4% of students did better than Joan, whereas 3.3% students did better than Betty.
  - Equivalently, we can just compare z-transforms (since the scores are assumed to be normal).

- Equivalently, we can just compare z-transforms (since the scores are assumed to be normal).

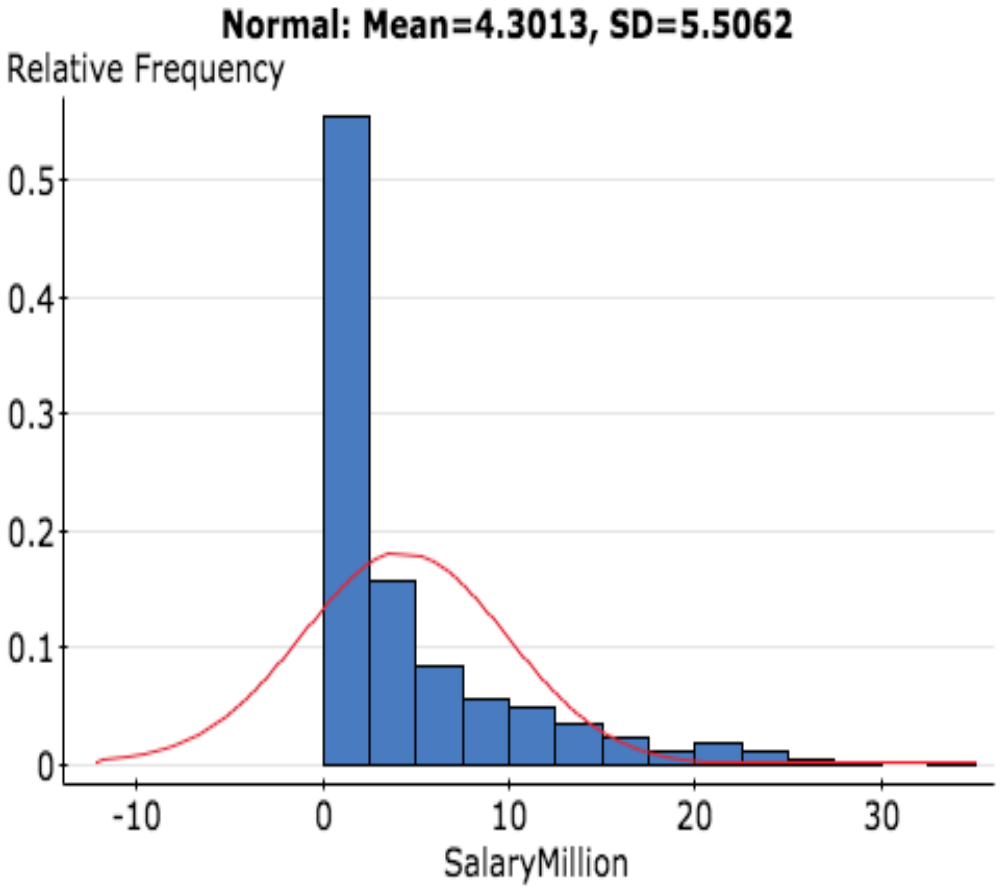
- We can also translate Joan's grade into a SAT grade using the z-transform.
- Since Joan is 2.2 standard deviations from the mean, this means if she took the SAT she would be **2.2 standard deviations** from the SAT mean.
- Thus Joan's ACT grade translated into a SAT grade

$$1025 + 2.2 \times 200 = 1465$$

- In other words, assuming that SAT and ACT grades were normally distribution. Joan's equivalent SAT grade is 1465.

Assessing the validity of the calculations?

- This is conceptually tricky.
- In all statistical analysis we need to take a step back and ask ourselves whether the calculations were **meaningful**.
- We are using the normal (or later the t-tables), this means we are assuming the z-score (or t-score) is normal (t-distributed).
- Is this always true?

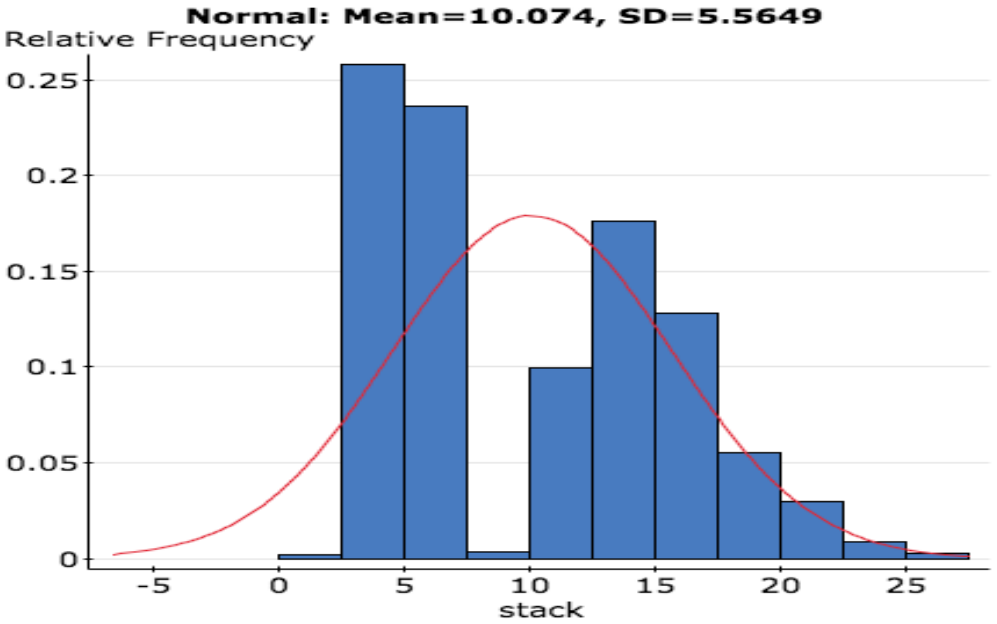


If baseball salaries are normally stributed 21% of baseball players would earn less than zero dollars in a year.

The calculation:

If baseball salaries were normal, what proportion of baseball players earn between 0 – 0.55 million dollars a year?

Question Time



Question: How well does the normal density approximate the proportions for this data?

A. The fit seems relatively good. Using the normal will give the correction proportions.
B. The normal will give the completely wrong answer for the proportion between 7.5 and 10.
C. The normal will give the wrong proportion for values less 7.5.
D. (B) and (C)

Question Time

- A farmer wants to enter either his cow or pig for the heaviest animal competition. The winning animal is the heaviest animal in its category (cows or pigs).
  - It is known that the weight of **cows** is approximately normally distributed with mean 280 pounds and standard deviation 20 pounds **(N(280,20)).**
  - The weight of **pigs** is approximately normally distributed with mean 250 pounds and standard deviation 50 pounds **(N(250,50)).**
  - His prize cow weighs 330 pounds and prize pig weighs 310 pounds.

  The contest only allows one animal per farmer, which animal should he enter?

A. His prize Cow
B. His prize Pig

Solution

- It makes sense to enter the heaviest animal is relative to its species.
- The z-score for the cow = (330-280)/20 = 2.5 standard deviations from the mean. This corresponds to the 99.3% percentile.
- The z-score for the pig is (310-250)/50 = 1.2 this corresponds to the 88.4 percentile. Despite the pig's weight lying further from the mean, there is a lot of variation in pig weight.
- The farmer should enter the cow, since only 0.7% of cows are heavier than her.

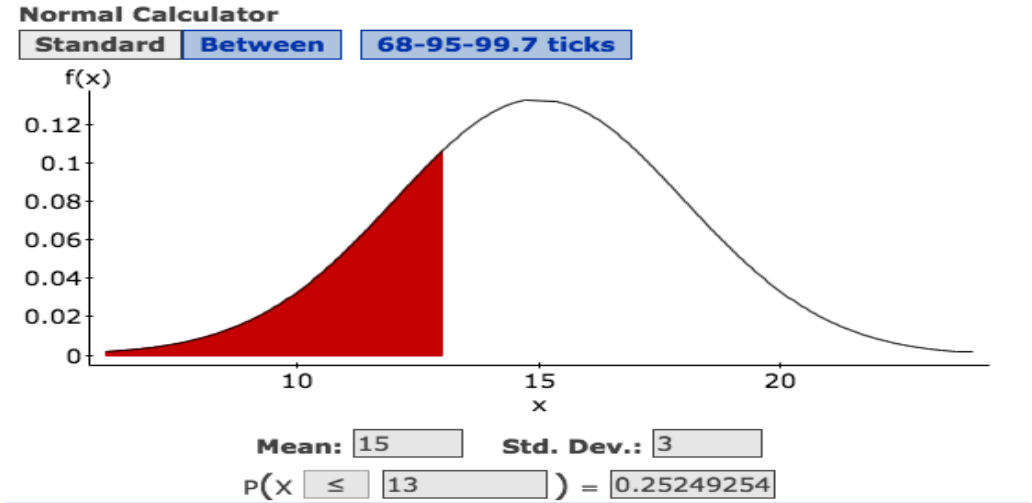
Question Time (difficult)

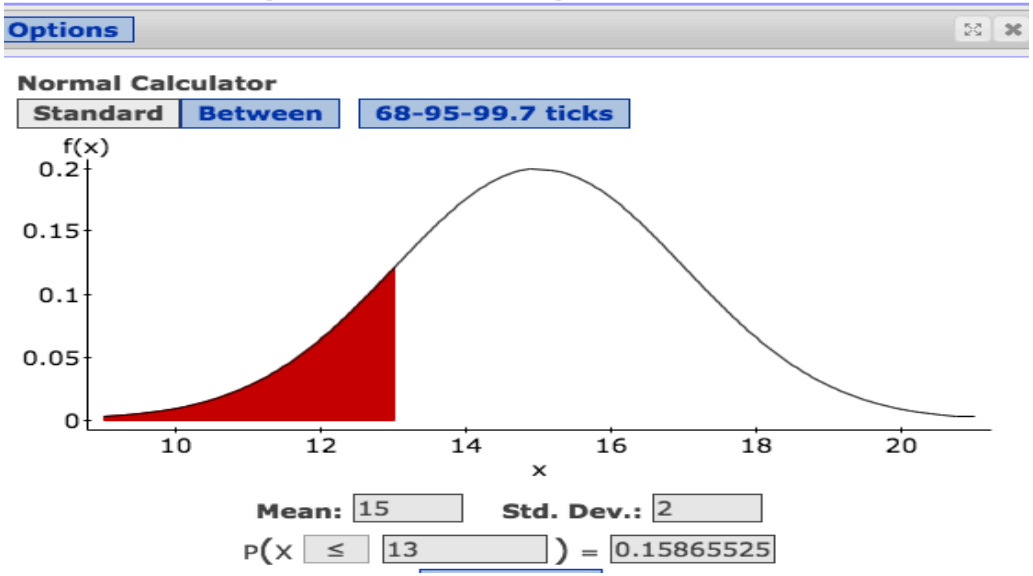
- It is important to monitor a foal in the first few days after birth. In particular it is important that the mare gives the foal essential colostrum. In order to do this, the foal must feed frequently from their mother (on average a newborn foal feeds 4 times in an hour). However, if the foal **feeds too frequently**, this suggests the mother is not providing enough colostrum and the foal may need veterinary assistance.
- The distribution of feed times (time between the feeds) of healthy foals follow a normal distribution. But for each breed of horses the distributions vary slightly:
  - Breed 1: N(15, 3) (mean 15 minutes, standard deviation 3)
  - Breed 2: N(15, 2) (mean 15 minutes, standard deviation 2)
- Suppose a vet uses a blanket threshold and examines any foal whose feed time drops less than 13 minutes. What are the implications of this policy?

A. Breed 1 will be examined **more** often than Breed 2.
B. Breed 2 will be examined **more** often than Breed 1.
C. Both breeds will be examined about the same.

Answer

- Each breed has its own distribution for healthy horses. But the same blanket threshold is used:





Because there is more variability (standard deviation) for breed one than breed two (but both have the same mean) we see that breed one will be examined more often than breed two.

This probabilities can be calculated using the z-score and looking but the normal tables.

Question Time (difficult)

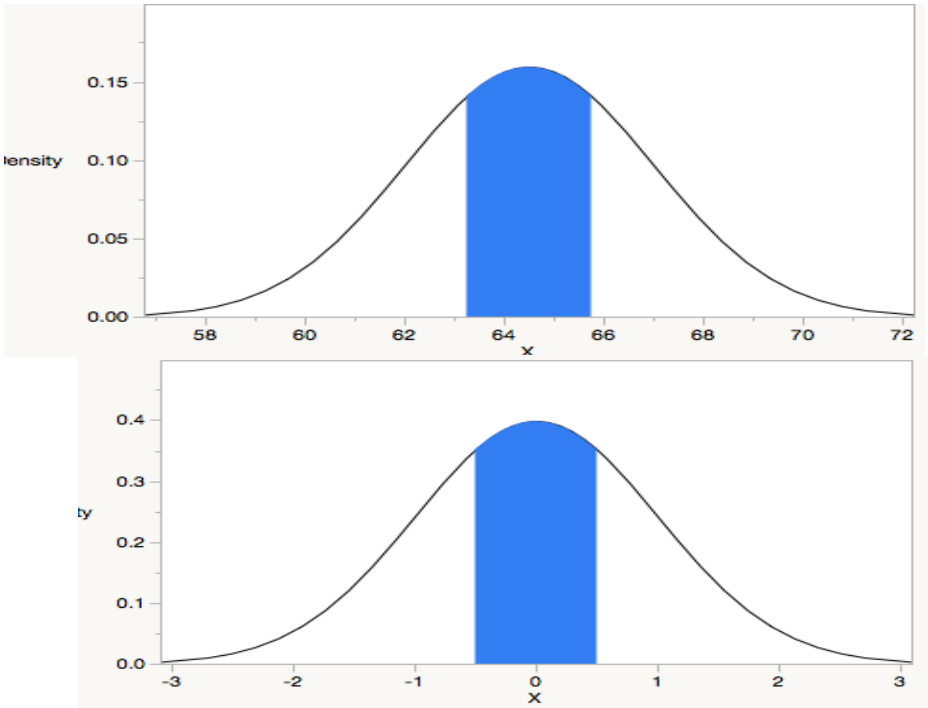
- Same question again but now we compare Breed 2 with Breed 3.
- The distribution of feed times (time between the feeds) of healthy foals follow a normal distribution. But for each breed of horses the distributions vary slightly:
  - Breed 2: N(15, 2) (mean 15 minutes, standard deviation 2)
  - Breed 3: N(16, 4) (mean 16 minutes, standard deviation 4)
- Suppose a vet uses a blanket threshold and examines any foal whose feed time drops less than 13 minutes. This means, in general:

A. Breed 2 will be examined **more** often than Breed 3.

B. Breed 3 will be examined **more** often than Breed 2.

Within z-standard deviation calculations

- **Question:** If a population is normally distributed what proportion of the population will lie within 0.5 standard deviations from the mean.
- **Answer:** This looks like mission impossible!
  - You may be asking, what is the mean, what is the standard deviation?? The point is that this information is not required.
  - When we say 0.5 standard deviations from the mean this is the same as the z-score being = -0.5 and 0.5.



Here we have two plots. One for the distribution of heights, which are assumed normal with mean 64.5 and standard deviation 2.5. The other is a *standard* normal with mean zero and standard deviation one.
The blue area corresponds to 0.5 standard deviations from the mean. The blue area for both plots is the same.

Within z-standard deviation calculations

- **Question:** If a population is normally distributed what proportion of the population will lie within 0.5 standard deviations from the mean.
- **Answer:** This looks like mission impossible!
  - You may be asking, what is the mean, what is the standard deviation?? This information is not required.
  - When we say 0.5 standard deviations from the mean this is the same as the **z-score is between** -0.5 and 0.5.
- Make a plot!

- The area between -0.5 and 0.5 is 0.6915 – 0.3085 = 0.383.
- About 38.3% of the population are **within 0.5 standard deviations** of the mean.

- Remember the number of standard deviations from the mean, corresponds to the z-score.

Question Time

- If a distribution is normally distributed what proportion of the population will lie within 1.5 standard deviations of the mean?

- A. 93.32%
- B. 6.68%
- C. 86.64%
- D. 1.5%

Inverse normal calculations

We may also want to find the observed range of values that correspond to a given proportion/ area under the curve.

For that, we use Table A **backward**:

- We first find the desired area/ proportion in the *body of the table*.
- We then read the corresponding *z-value* from the left column and top row.

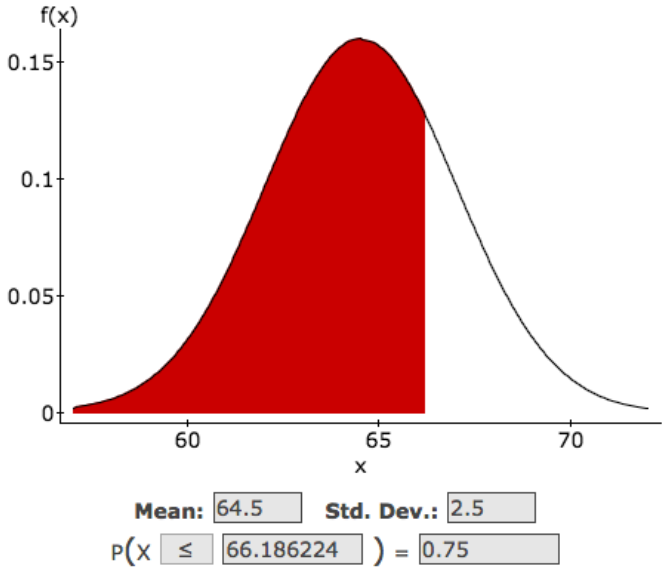
TABLE A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143

For an area of 1.25% (0.0125) to the left of *z*, the *z*-value is −2.24.

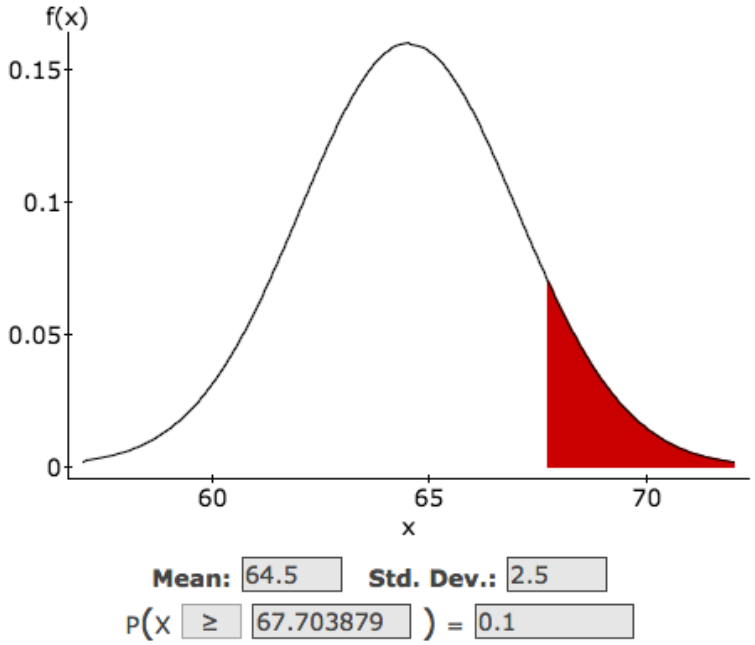
Example: Female heights

- Female heights tend to be normally distributed with N(64.5,2.5).
- Questions:
  - (a) How tall is a female in the 75% percentile?
- Answers:
  - (a) Look up 0.75 inside the z-table, it is 0.674. This means that someone who is in the 75 percentile is 0.674 standard deviations to the right of the mean. That person is $64.5 + 0.674 \times 2.5 = 66.2$ inches tall.



Example: Female heights

- Female heights tend to be normally distributed with N(64.5,2.5).
- Questions:
  - (b) How tall is a female who is in the top 10% percentile?
- Answers:
  - (b) Top 10% = 90% percentile. Look up 0.9 in the z-table, it is 1.28. Using the same argument as above that person is 64.5 + 1.28 × 2.5 = 67.7 inches tall.



Question Time

- Female heights tend to be normally distributed with N(64.5,2.5) (mean 64.5 inches and standard deviation 2.5 inches). How tall is a female in the top 20^{th} percentile for heights?

  A. 66.63 inches

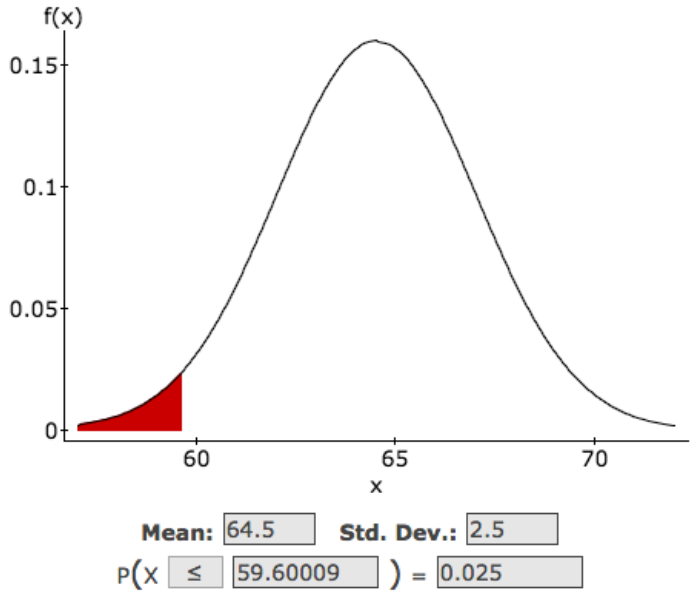
  B. 66.5 inches

  C. 62.38 inches

  D. 65 inches

Example: Female heights

- Female heights tend to be normally distributed with N(64.5,2.5).
- Questions:
  - (c) How tall is a female who is in the bottom 2.5% percentile?
- Answers:
  - (c) Look up 0.025 in the z-tables – 1.96. Using the same argument as above that person is 64.5 -1.96 × 2.5 = 59.6 inches tall.



- **Question**: Construct an interval **centered about the mean**, where 95% of female heights lie.
  - Answer: Look up 2.5% and 97.5% in z-tables; [-1.96, 1.96] (this interval is symmetric about 0).
  - This interval stays that 95% of heights will lie within 1.96 standard deviations (either way) of the mean.
  - Translating this into heights we have that 95% of heights lie between

    $[64.5 - 1.96 \times 2.5,\ 64.5 + 1.96 \times 2.5] = [59.6, 69.4]$ inches.

Question Time

- Female heights tend to be normally distributed with mean 64 inches and standard deviation 2.5: N(64.5,2.5). Construct an interval **centered** about the mean of 64.5 where 80% of heights will lie.

  A. [61.3, 67.7]

  B. 66.63

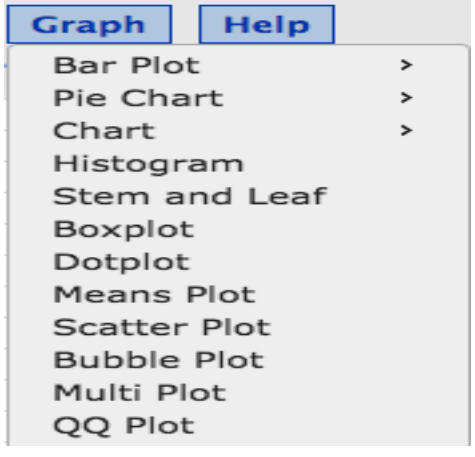
  C. [62.38, 66.63]

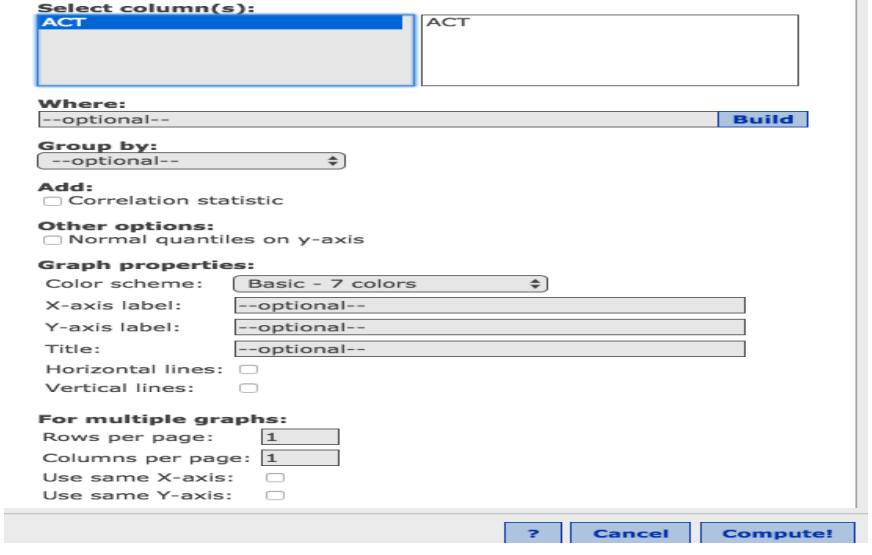
Topic: QQplots

- Learning Targets:
  - Understand that QQplots are a graphical tool that allows to see if the data has approximately a normal density curve.
  - Understand how deviations from the straight line explain features about the underlying true distribution.

Using QQplots to check normality of data

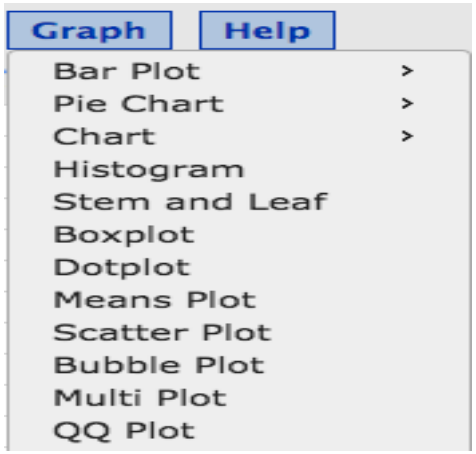
- By simply superimposing a normal distribution over a histogram it is difficult to check how close the distribution is to normal.
- Typically to check for normality of data we make a QQplot.
- The idea is behind this plot is similar to checking the 68-95-99.7% but extended to all multiples of the standard deviation not just 1,2, and 3.
- The data is close to normal if the points lie along the x=y line.
- Warning: The QQplot has **nothing** to do with linear regression. The line that you see in the plot is **not** the line of best fit.
- In the following few slides we consider a few examples:
- Making a QQplot in Statcrunch.

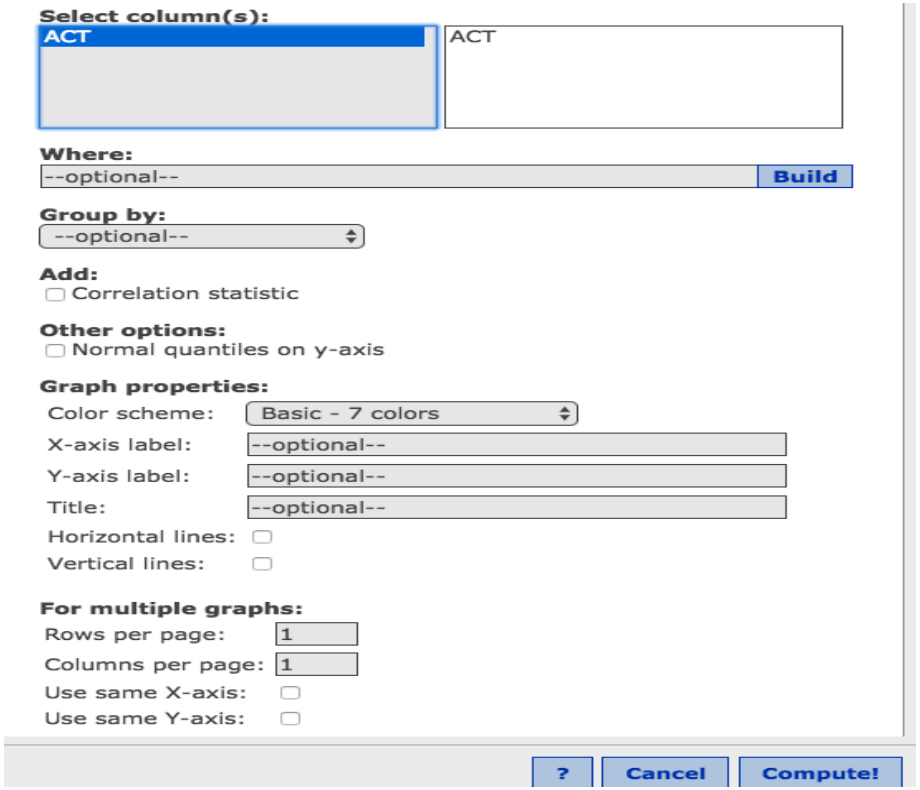




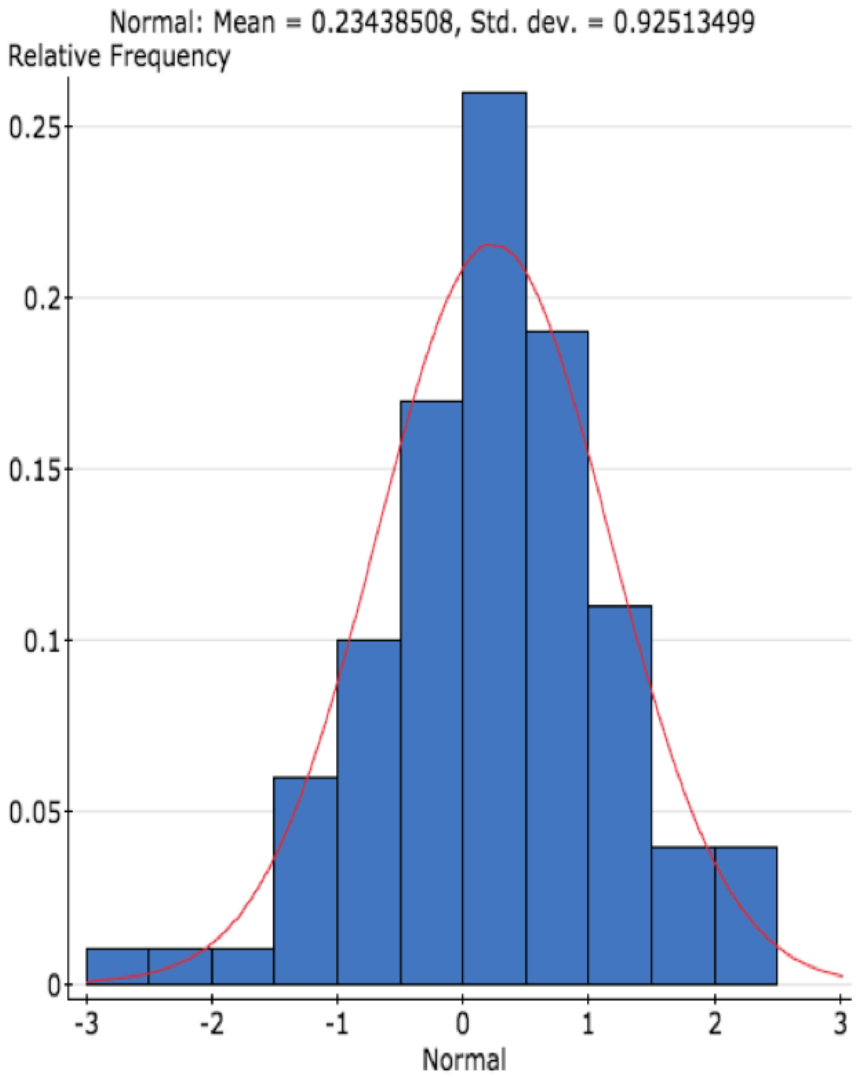
Using QQplots to check normality of data

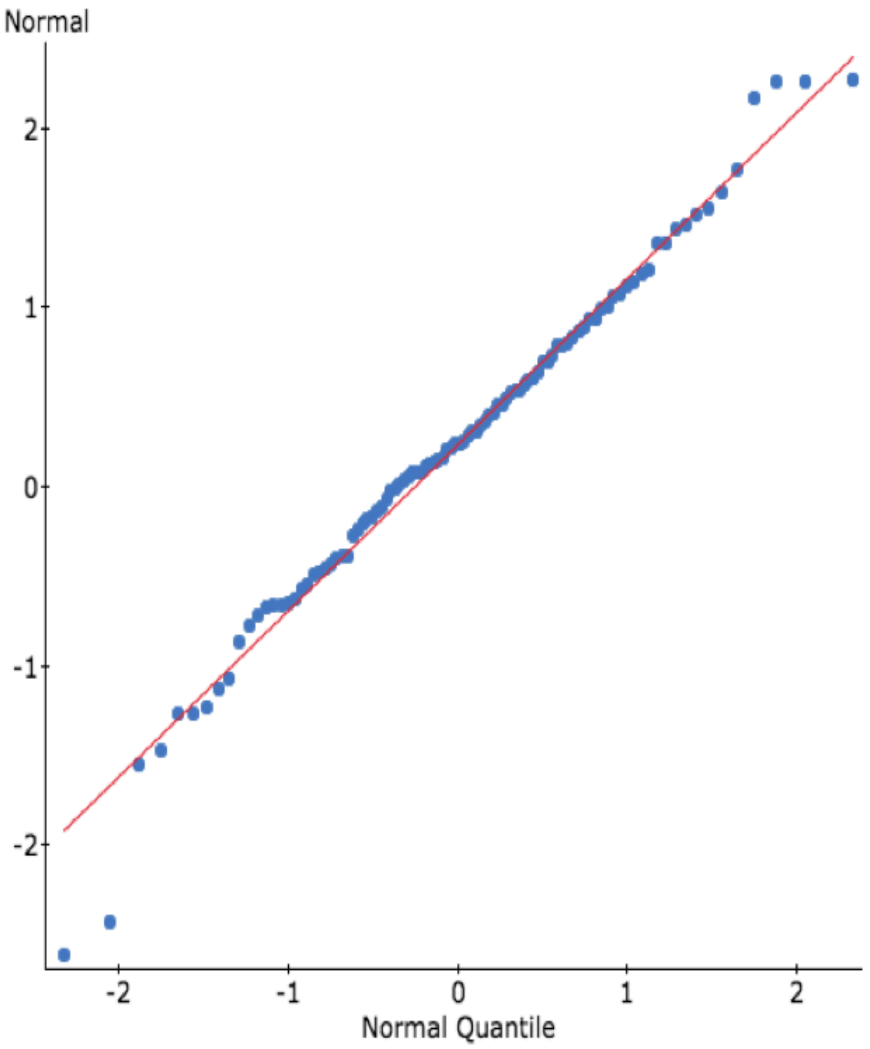
- Making a QQplot in Statcrunch.





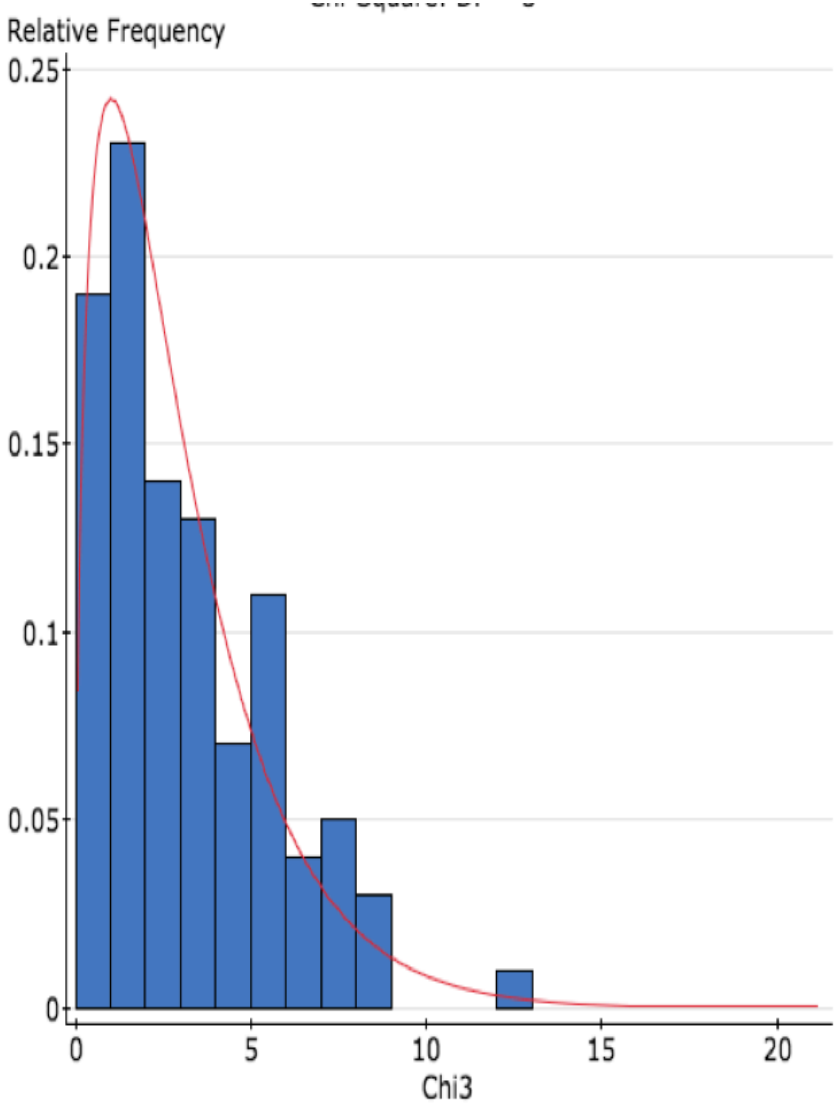
QQplot of normal data

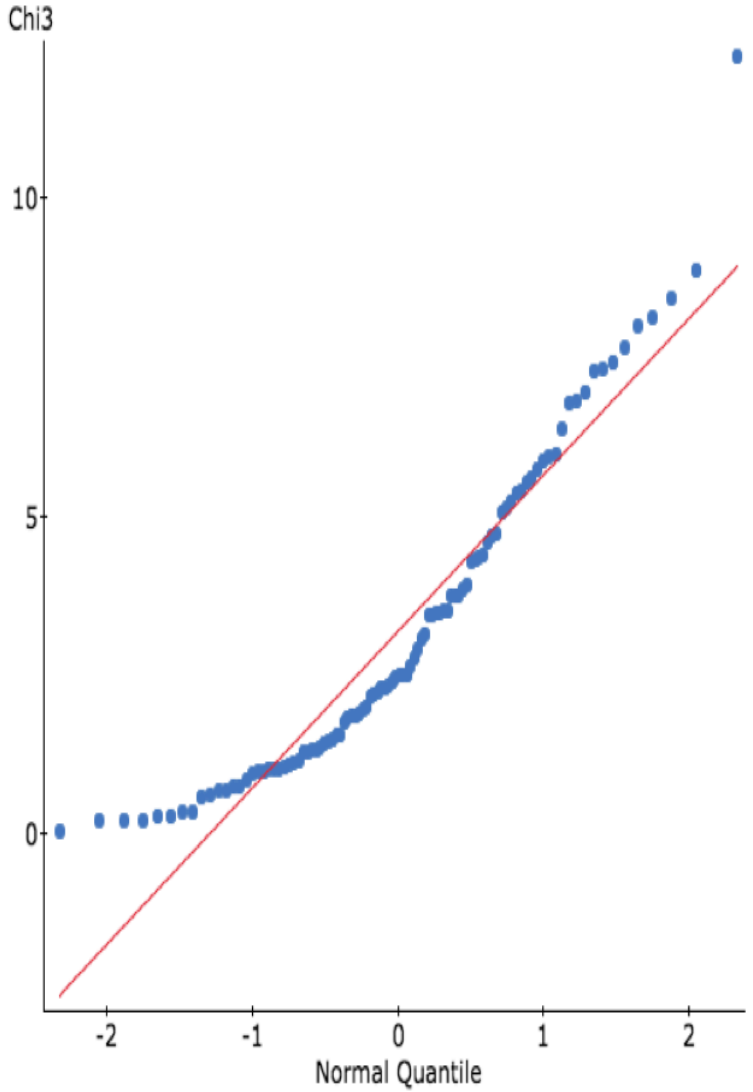




Observe that most points lie close to x=y line. Few in the tails lie off the line.

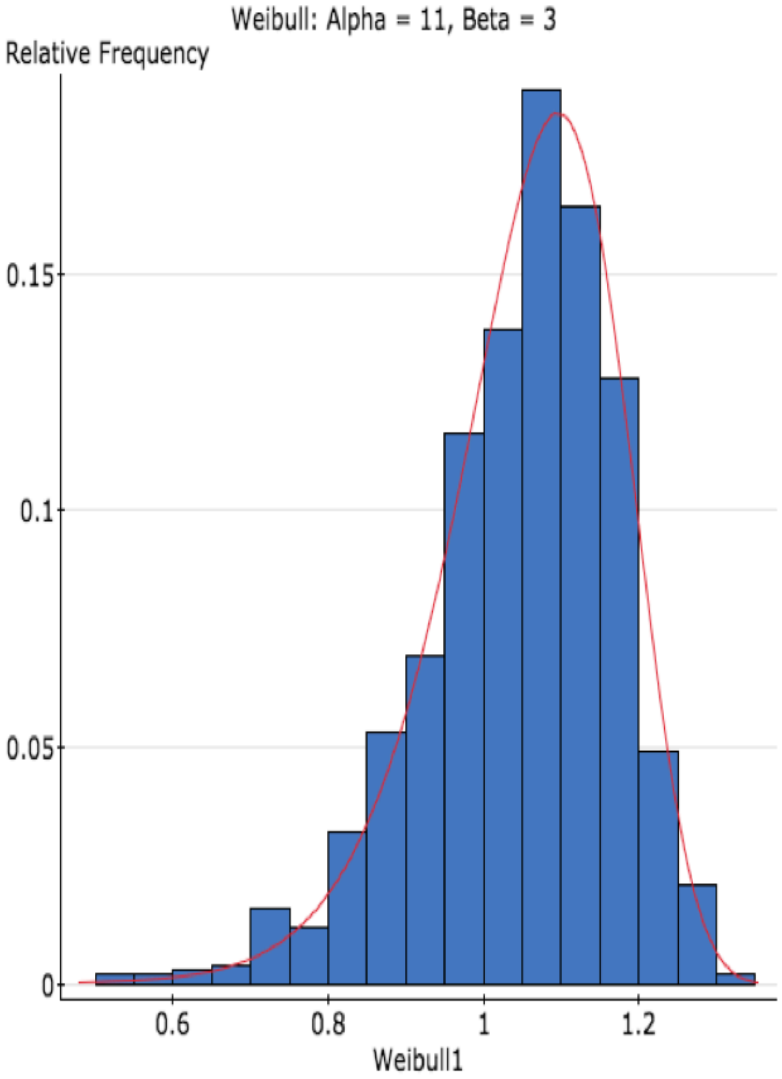
QQplot of right skewed data

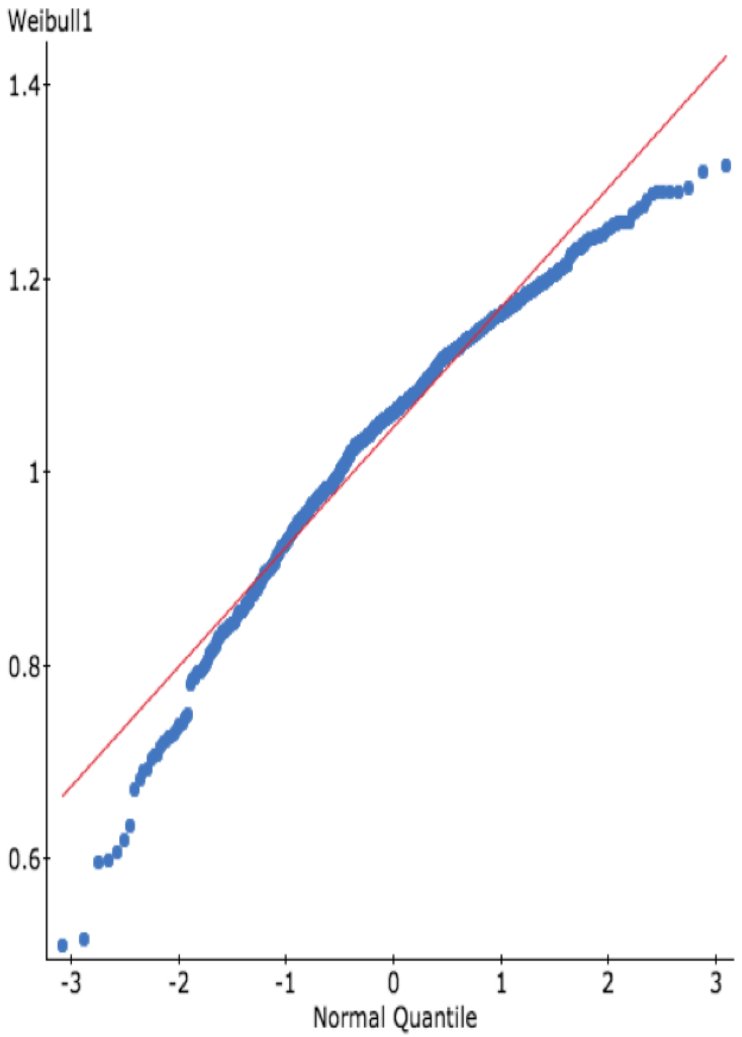




For right skewed data the Qqplot has a U-type shape

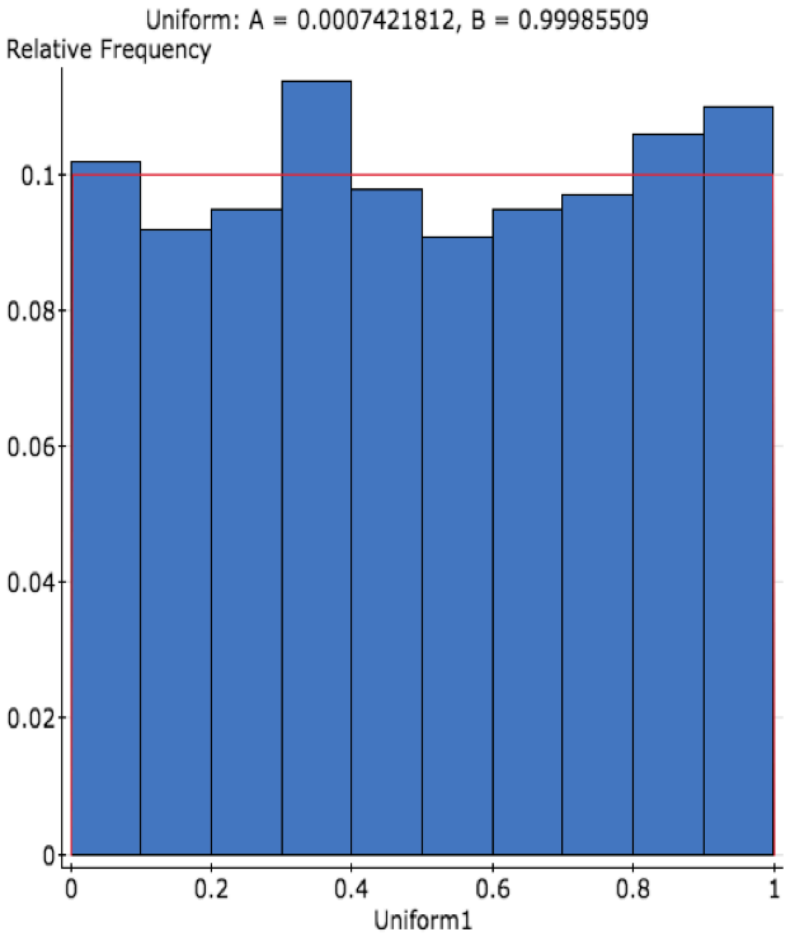
QQplot for left skewed data

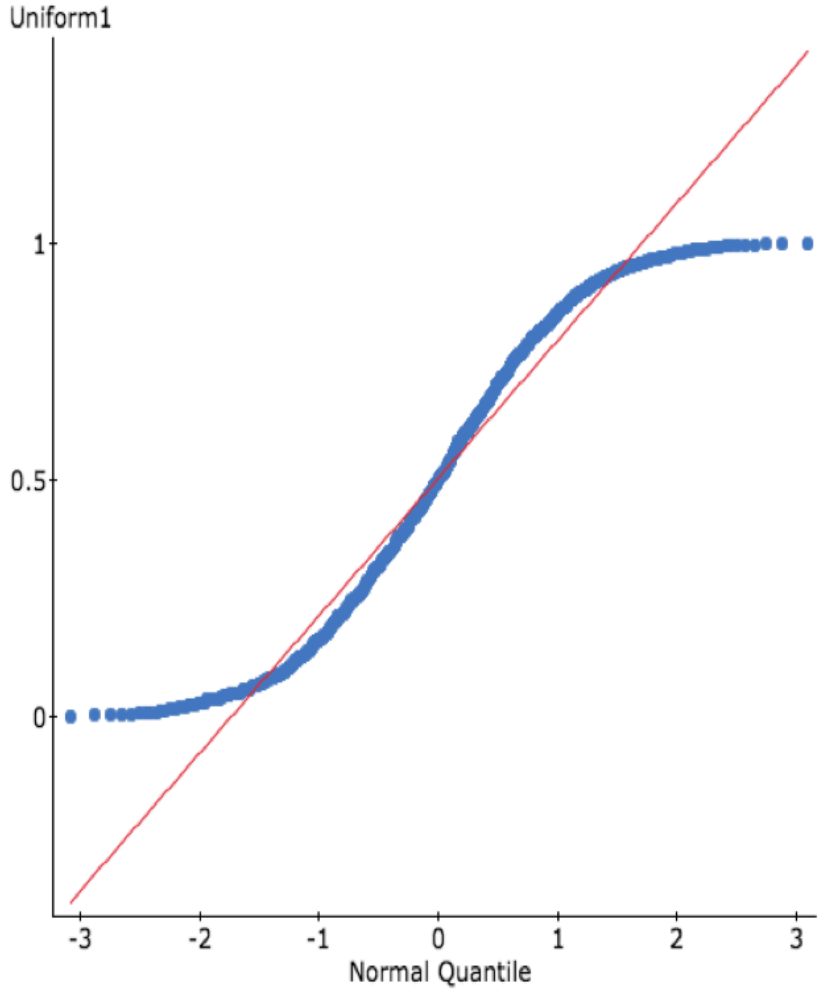




QQplot for left skewed data looks like an inverted U

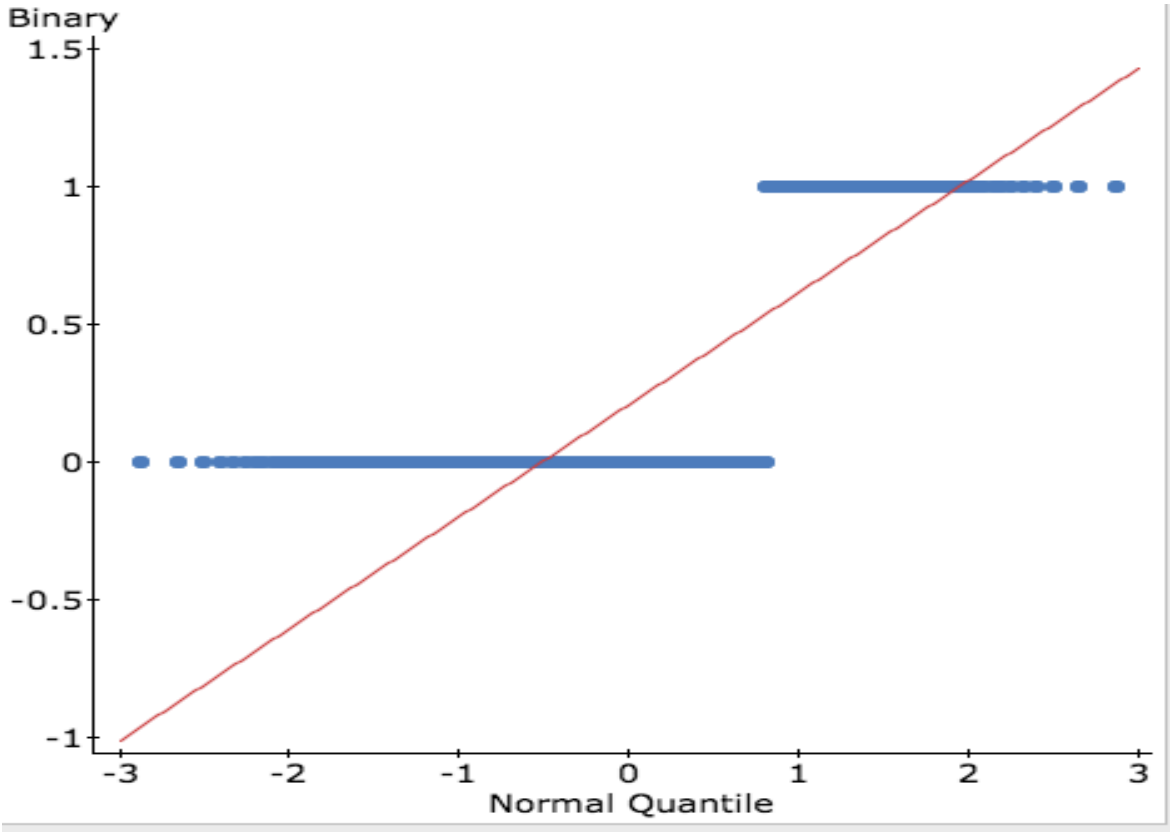
QQplot for uniform data





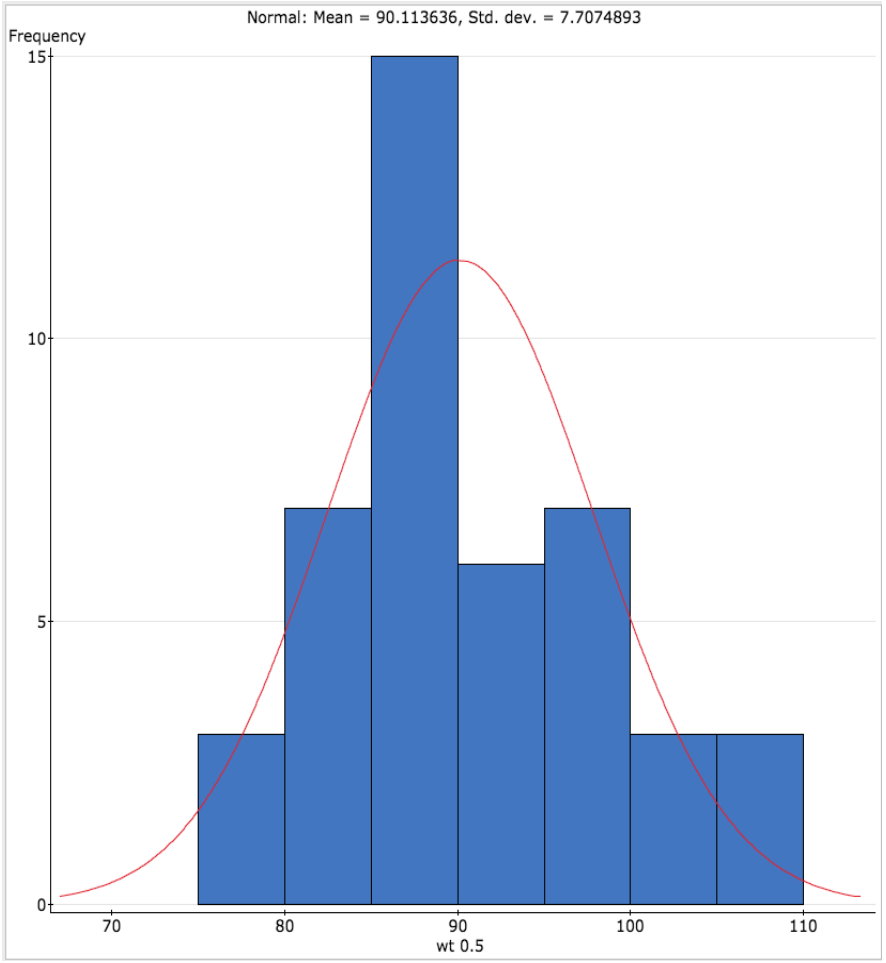
QQplot for uniform and thick tailed data (data whose tails are not much thinner than the center) have an S shape.

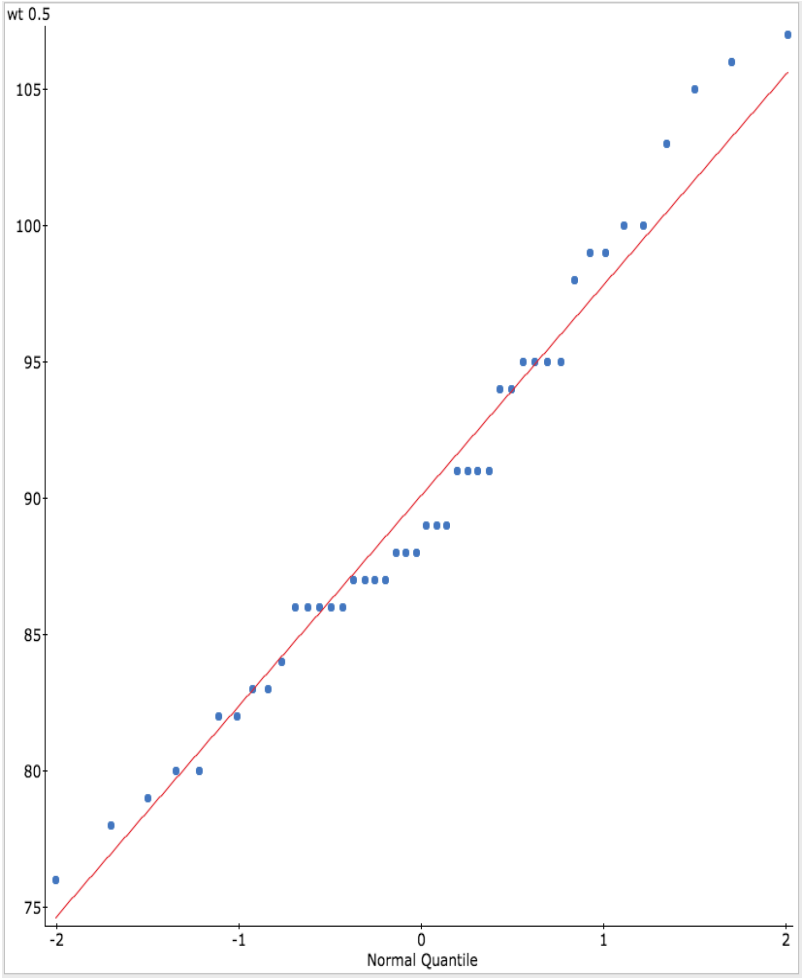
QQplot for binary response data



In this data set, the response is either 0 or 1. The vertical lines correspond to each of these responses.

QQplot of calf data

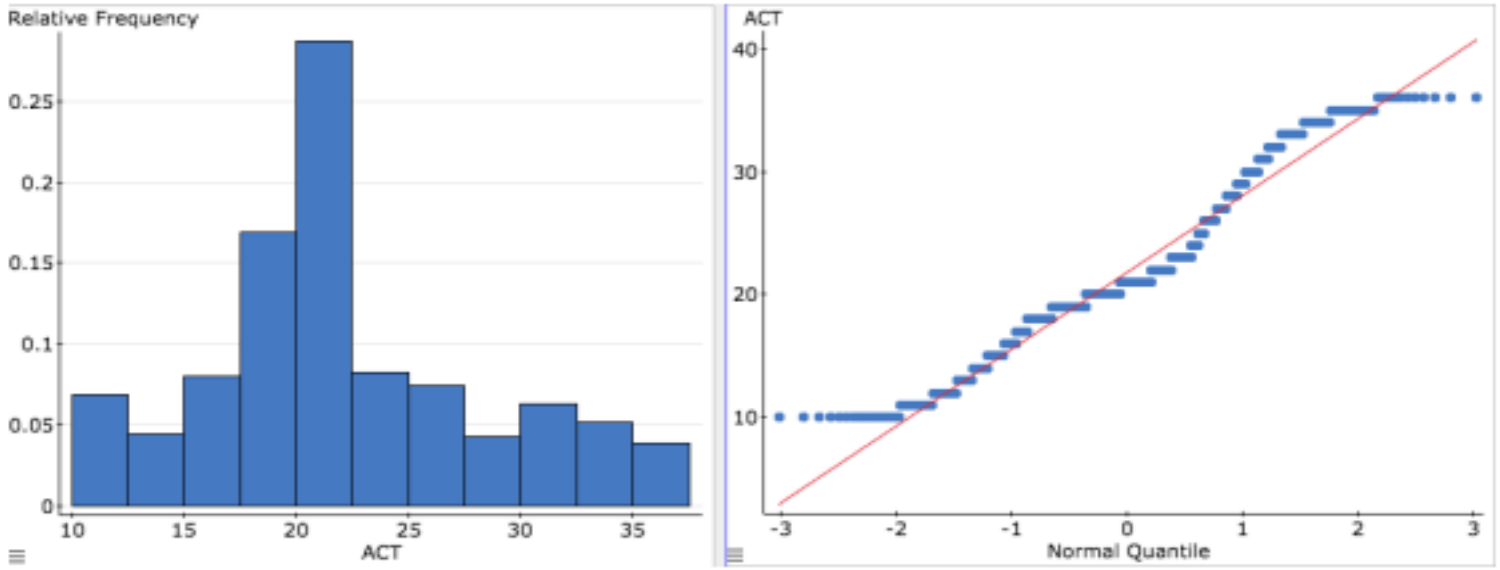




The horizontal lines we see are due to several weights having the same value (due to rounding). Eg. The first horizontal line corresponds to 5 calves with the same weight. The weights are not exactly normal, but it does not deviate massively from normality.

Question Time

- Based on the plot below which statement is correct



- (A) The horizontal lines show that the ACT scores are integer valued
- (B) The distribution of ACT scores differ from normality especially in the right and left tails.
- (C) The ACT scores lie close to the line and are normal.
- (D) There is a clear linear correlation in the ACT scores.
- (E) A and B.
- (F) C and D.

Accompanying problems associated with this Chapter

- HW 4
- HW 5