Statistics for Computing

MA4413 Lecture 3A

Kevin O'Brien

Kevin.obrien@ul.ie

Dept. of Mathematics & Statistics,
University *of* Limerick

Autumn Semester 2013

Today's Class

- More on Graphical methods
  - Bar charts
  - Box-and-whisker plots
- Discrete probability distributions
  - Binomial Experiments
  - The Binomial Probability distribution

Bar plots

- A bar plot displays the frequency (or relative frequency) for all observations of a discrete random variable.
- A bar plot is much like a histogram, in that the heights of columns represent the frequency (or relative frequency) of each outcome.
- Each outcome of a random experiment corresponds to one and only one column of the bar plot.
- A bar plot differs from a histogram in that the columns are distinct and separated from each other by a small distance.

Bar plots

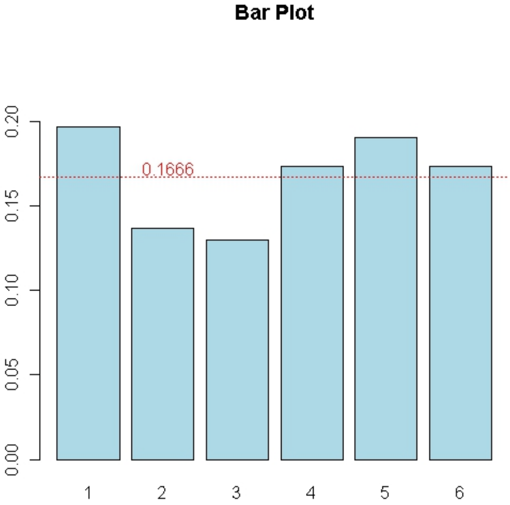
Suppose we roll a die 300 times, and obtain the following results

Outcome	1	2	3	4	5	6
Frequency	59	41	39	52	57	52
Rel. Freq	0.196	0.136	0.130	0.173	0.190	0.173

On the next slide is the bar plot of the relative frequencies of the outcomes of die throw experiment. Included on the bar plot is the theoretical probability of each outcome. As each outcome is equally probable, this is just a straight line.

Minor deviations from the theoretical probability can often be assumed to be as a result of random error. In the case of large deviations, there may be a flawed assumption about the theoretical probabilities.

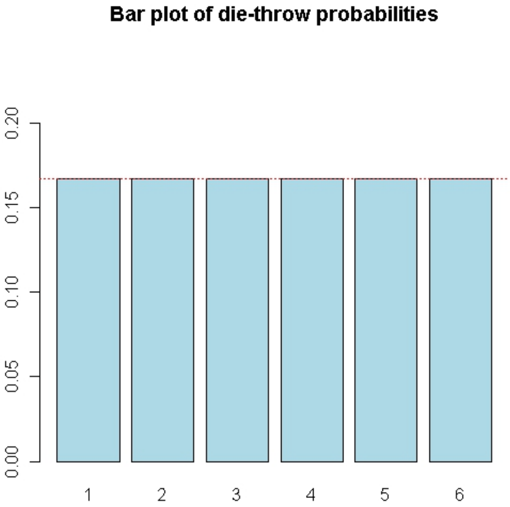
Relative Frequency Bar Plots



Relative Frequency Bar Plots

- Just as bar plots can be used to graphically depict observed relative frequencies, they can be used to depict the theoretical probabilities of each outcome.
- We will be using bar plots to visualize the theoretical probabilities of outcomes of discrete random variables.
- For this module, bar plots are assumed to be used for this purpose, unless it is clearly expressed otherwise.
- On the next slide is a bar plot of the probabilities of each outcome of a dice throw.

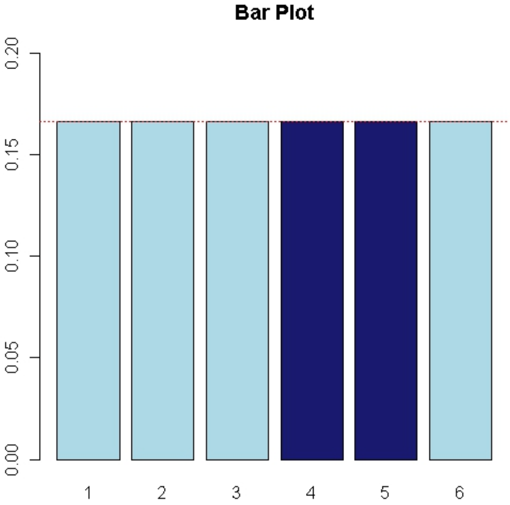
Bar Plots


Bar plot of die-throw probabilities
0.00
0.05
0.10
0.15
0.20
1
2
3
4
5
6

Bar Plots

- Bar plots are useful in that they visualize ‘events’.
- Consider the event where either a ‘4’ or a ‘5’ is thrown.
- The relevant columns for this event are shaded (next slide).
- We will be using bar plots for depicting specific events in upcoming material

Bar Plots



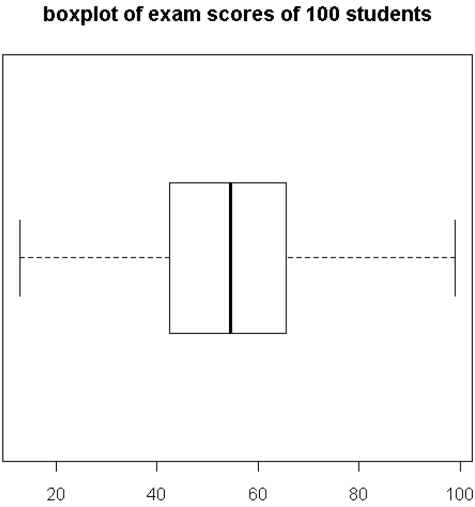
Boxplots

- The second graphical method we will be looking at today is the 'box-and-whisker' plot (commonly just referred to as 'boxplots')
- The boxplots is a useful tool for assessing the distribution of a dataset, by means of a visual summary.
- Recall the data set of the exam scores of 100 students from yesterday's class (see next slide).
- The quartiles of the data set were $Q_1 = 42.5$, $Q_2 = 54.5$ (with Q_2 being the median), and $Q_3 = 65.5$ respectively.
- The interquartile range is $Q_3 - Q_1 = 23$
- The boxplot of the distribution is featured on the next slide.

Table: Exam results of 100 students

13	21	22	23	24	25	26	28	29	30
31	32	33	34	35	36	36	36	37	38
39	41	41	41	42	43	44	44	44	45
45	46	47	49	50	51	51	52	53	53
53	53	53	54	54	54	54	54	54	54
55	55	55	56	56	56	57	57	58	59
62	63	63	63	63	64	64	64	64	64
65	65	65	65	65	66	66	66	67	69
71	71	72	72	73	74	75	76	76	76
77	82	84	85	87	88	91	91	92	99

Boxplots



Boxplots

- The boxplot is a visual summary containing important aspects of a distribution.
- The main component of the plot , the ‘***box***’, stretches from the ***lower hinge***, defined as Q_1, to the ***upperhinge***, defined as Q_3 .
- The median is shown as a line across the box.
- Therefore the box contains the middle half of the scores in the distribution.
- 1/4 of the distribution is between the median line and the upper hinge. Similary 1/4 of the distribution is between the median line and the lower hinge.

Boxplots

- On either side of the box are the ***whiskers***.
- To find where to place the whiskers, we must first compute the location of the ***fences***, and determine whether or not there are any ***outliers*** present.
- Firstly, we must compute the location of the ***lower fence***.

$$\text{Lower Fence} = Q_1 - 1.5 \times IQR$$

- For our example, the lower fence is

$$\text{Lower Fence} = 42.5 - 1.5 \times 23 = 42.5 - 34.5 = 8$$

Boxplots

- The lower fence is used to determine whether there are any outliers in the lower half of the data set.
- If there is any observed value less than the lower fence, it is considered an outlier.
- The first whisker is drawn at the location of the lowest value that is not considered an outlier.
- If no values are considered outliers, then the whisker is drawn at the location of the smallest value of the dataset.
- For our dataset, the lowest value is 13, which is not less than the lower fence.
- Therefore we draw the first whisker , a vertical line, at this location.
- A horizontal line is drawn connecting the location of this whisker to Q_1.

Boxplots

- Any value considered to be an outlier should be indicated with an asterisk or a small circle.
- We will see an example of a boxplot with outliers in due course.

Boxplots

- Now we must compute the location of the ***upper fence***.

$$\text{Upper Fence} = Q_3 + 1.5 \times IQR$$

- For our example, the upper fence is

$$\text{Upper Fence} = 65.5 + 1.5 \times 23 = 65.5 + 34.5 = 100$$

Boxplots

- The upper fence is used to determine whether there are any outliers in the upper half of the data set.
- If there is any observed value greater than the upper fence, it is considered an outlier.
- The second whisker is drawn at the location of the highest value that is not considered an outlier.
- If no values are considered outliers, then the whisker is drawn at the location of the highest value of the dataset.
- For our dataset, the highest value is 99, which is less than the upper fence.
- Therefore we draw the second whisker , a vertical line, at this location.
- A horizontal line is drawn connecting the location of this whisker to Q_3.

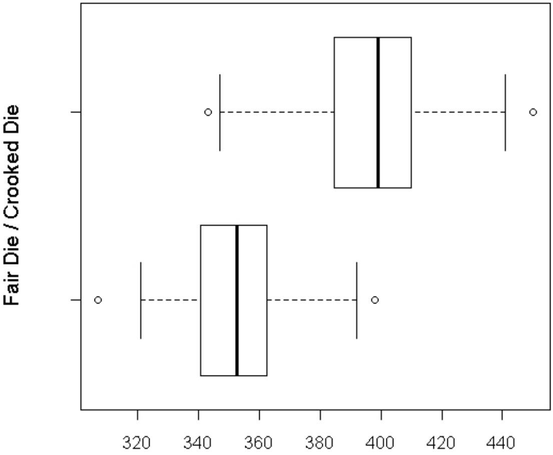
Boxplots

- Remark: If you do not get a sensible value for either the upper or lower fence, you can replace it with the nearest sensible value
- For example, suppose we got a negative lower fence value. It does not make sense to get a negative score in an exam.
- In this case, we could replace the value with a value of 0.
- similarly for the upper fence: any fence value greater than 100 should be replaced with the value of 100.

Boxplots

- Boxplots are very useful in comparing the distributions of two or more data sets.
- Recall the experiment of 60 students, each throwing a die 100 times.
- Suppose they perform this experiment twice, firstly with a fair die, and then with a crooked die.
- (The probability of the outcomes from the crooked die are as per yesterday's class).
- Boxplots can use used to compare the distribution of the outcomes of both experiments.

Boxplots



Discrete Probability Distributions

- Over the next set of lectures, we are now going to look at two important discrete probability distributions
- The first is the ***binomial*** probability distribution.
- The second is the Poisson probability distribution.
- In `R`, calculations are performed using the `binom` family of functions and `pois` family of functions respectively.

Binomial Experiment

A binomial experiment (also known as a Bernoulli trial) is a statistical experiment that has the following properties:

- The experiment consists of n repeated trials.
- Each trial can result in just two possible outcomes. We call one of these outcomes a ***success*** and the other, a ***failure***.
- The probability of success, denoted by p, is the same on every trial.
- The trials are independent; that is, the outcome on one trial does not affect the outcome on other trials.

Binomial Experiment

Consider the following statistical experiment. You flip a coin five times and count the number of times the coin lands on heads. This is a binomial experiment because:

- The experiment consists of repeated trials. We flip a coin five times.
- Each trial can result in just two possible outcomes : heads or tails.
- The probability of success is constant : 0.5 on every trial.
- The trials are independent; that is, getting heads on one trial does not affect whether we get heads on other trials.

Binomial Probability

- A binomial experiment with n trials and probability p of success will be denoted by
$$B(n,p)$$
- Frequently, we are interested in the ***number of successes*** in a binomial experiment, not in the order in which they occur.
- Furthermore, we are interested in the probability of that number of successes.

Binomial Probability

The probability of exactly k successes in a binomial experiment B(n, p) is given by

$$P(X = k) = P(k \text{ successes }) = {}^{n}C_{k} \times p^{k} \times (1-p)^{n-k}$$

- X: Discrete random variable for the number of successes (variable name)
- k : Number of successes (numeric value)
  - $P(X = k)$ "probability that the number of success is k".
- n : number of independent trials
- p : probability of a success in any of the n trial.
- $1 - p$: probability of a failure in any of the n trial.

Binomial Example

Suppose a die is tossed 5 times. What is the probability of getting exactly 2 fours?

Solution:

This is a binomial experiment in which

- a success is defined as an outcome of ‘4’.
- the number of trials is equal to $n = 5$,
- the number of successes is equal to $k = 2$,
- the number of failures is equal to 3,
- the probability of success on a single trial is 1/6,
- the probability of failure on a single trial is 5/6.

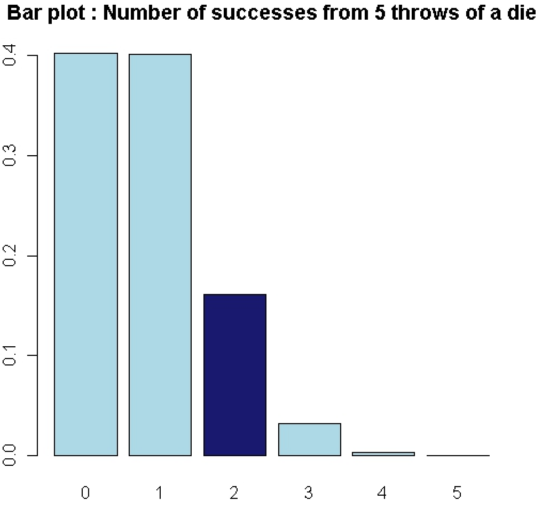
Binomial Example

Therefore, the probability of getting exactly 2 fours is:

$$P(X=2) =^5 C_2 \times (1/6)^2 \times (5/6)^3 = 0.161$$

Remark: ${}^5C_2 = 10$

Binomial Example

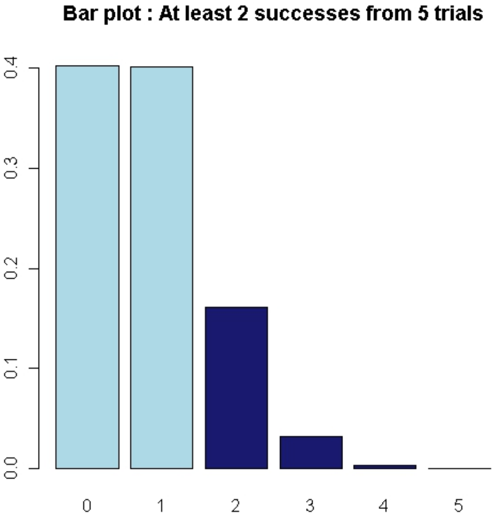


Binomial Probability

Remark : The sum of the probabilities of each of the possible outcomes (i.e. no fours, one four etc) is equal to one.

$$P(X=0)+P(X=1)+\ldots+P(X=5)=1$$

Binomial Example: At least two successes



Binomial Example: At least two successes

- Suppose we were asked to find the probability of ***at least*** 2 fours.
- Can you suggest the most efficient way of computing this?
- Suggestion: Compute $P(X=0)$ and $P(X=1)$.
- Together these probabilities are the complement probability of what we require.
- $P(X \geq 2) = 1-(P(X=0)+P(X=1))$.
- (We will continue with this in future classes).

Cumulative Distribution Function

The cumulative distribution function (c.d.f.) of a discrete random variable X is the function $F(t)$ which tells you the probability that X is less than or equal to t.
So if X has p.d.f. P(X = x), we have:

$$F(t) = P(X \leq t) = \sum_{(i=0)}^{(i=t)} P(X = x)$$

In other words, for each value that X can be which is less than or equal to t, work out the probability that X is that value and add up all such results.

Binomial Example: Sample Problem

Suppose there are twelve multiple choice questions in an English class quiz. Each question has five possible answers, and only one of them is correct. Find the probability of having four or less correct answers if a student attempts to answer every question at random.

Solution: Since only one out of five possible answers is correct, the probability of answering a question correctly by random is $1/5 = 0.2$. We can find the probability of having exactly 4 correct answers by random attempts as follows.(Blackboard. Correct Answer is 13.29%)

```
> dbinom(4, size=12, prob=0.2)
[1] 0.1329
```