Basic Statistics

Mo D Jabeen

April 4, 2023

The example used will be the size of carrots in a farm, on the current batch.

1 General

1.1 What is population and sample?

Population: Entire data set, ie all the carrots.
Sample: A portion of the dataset, ie 30 carrots.

1.2 What is a parameter or statistics?

Parameter: Attribute from population ie mean of carrots based on all carrots.
Statistic: Attribute from sample ie mean from 30 carrots.

Random selection of the sample allows for much better statistical inference as it removes any bias.

1.3 What is a regression test?

Determines if a prediction variables changes effect the outcome variable.

1.4 What are degrees of freedom?

The number of independent pieces of info used to calculate a statistic.

1.5 What is the mean, expectation and standard deviation?

The mean is the frequency of each value occurring, multiplied by the value all summed for each random variable.

$$\overline{x} = 1/n(\Sigma f x) \tag{1.1}$$

The expectation is the probability of each value multiplied by the value, summed for all values. This is the value the mean tends to as the sample size increases.

$$E(x) = \Sigma x P(X = x) \tag{1.2}$$

The variance is the average of the squared difference from the mean.

$$\sigma^2 = E(X^2) - (E(X))^2 \tag{1.3}$$

The standard deviation is the square root of the variance, showing essentially the average distance from the mean:σ.

An overall shift to all data points will effect expectation and not variance:

$$E(X \pm a) = E(X) \pm a \tag{1.4}$$

$$Var(X \pm a) = Var(X) \tag{1.5}$$

An overall multiplier to all data points effects both expectation and variance:

$$E(aX) = aE(X) \tag{1.6}$$

$$Var(aX) = a^2 Var(X) \tag{1.7}$$

1.6 Geometric Mean

Mean based on the product of all values, finding the nth root.

$$\Pi x^{1/N} \tag{1.8}$$

N is the number of values.

1.6.1 What is it good for?

- For fraction based values ie percentages
- Dependant values
- Wildly varying values, less skewed by large data values

2 Continuos variables

2.1 What is the difference between continuos and discrete variables?

Discrete variables are a known list of possible numbers
Continuos random variables are infinite.

2.2 What is relative frequency density and how does it translate to probability?

The relative frequency density; is a measurement of the relative frequency over a class width (interval between two values). Relative frequency is how many times something happens between two values compared to number of measurements, class width the measurement period.

$$\frac{Relative frequency}{class width} = Relative frequency density \tag{2.1}$$

The probability density function f(x); is the relative frequency density as n increases and the class width decreases.

Area under a plotted f(x) gives the probability for that range of continuos variables.

$$P(X < x) = \int_{-\infty}^{x} f(x) \tag{2.2}$$

$$\frac{d}{dx} P(X < x) = f(x) \tag{2.3}$$

2.3 How do you calculate the Median?

The median value (m) is when the probability for values above and below are 0.5.

$$\int_{m}^{\inf} f(x) = 0.5 \tag{2.4}$$

2.4 How do you calculate a Percentile?

The Xth percentile is the value below which the probability is X/100:
90th percentile

$$P(X < x_{90th}) = 0.9 \tag{2.5}$$

2.5 How do you calculate the Mean?

If assume the a small width of delta x, the mean will be $\sum x(f(x)\delta(x))$ (the brackets give the probability and multiplying by x gives the mean). As delta x tends to 0 this becomes:

$$\bar{x} = \int x f(x) \tag{2.6}$$

The population mean is then :

$$\mu = \int_{-\infty}^{\infty} x f(x) dx \tag{2.7}$$

The variance is:

$$Var(x) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx \tag{2.8}$$

2.6 Inferential Statistics

Draw conclusions and predictions based on the data.

2.6.1 What is a confidence interval ?

Taking into account sampling error give a range of values for the parameter and certainty percentage.
i.e. Carrots mean is [8 12cm] at 95%.

3 Data Validation

- Constraints
- Visual graphing
- Distribution measurements (check if mean, median and mode are similar)
- Good fit tests (Chi squared)

- Independence test (Chi squared)
- Check for missing or errored data

3.1 Skew

A dist can have right(positive), or left(negative) or zero skew. A quick check for skew is a frequency histogram.
Right skew, has a very long tail, on its right side (mean > median).

Normal dist has 0 skew, all symmetry dists have 0 skew.

$$Pearsons\ Skew = 3 * \frac{Mean - Median}{std} \tag{3.1}$$

If skewed you can transform using square rooting, ln, log10 respectively as the skew is progressively worse.

3.2 Kurtosis

Measure of taildness of a distribution, how often outliers occur. Measured by checking the standardized 4th moment of a distribution.

$$Kurtosis = \hat{\mu_4} = \frac{\mu_4}{\sigma^4} \tag{3.2}$$

μ_4 is the unstandardized 4th moment.
The above formula is biased, unbiased formula uses a correction of sample size (find online)

3.3 Effect Size

Practical significance measure of if the found relationship has a real world impact.
Both Cohen d and pearsons R have guides on values showing levels of significance.

3.3.1 What is Cohens d?

Size of the difference between two groups.

$$d = \frac{\bar{x_1} - \bar{x_2}}{s} \tag{3.3}$$

s is the std from either a pooled data of the two groups or a control group.

3.3.2 What is Pearsons r?

Measures correlation by measuring variability, ie measure the effect size nutrients in the soil changes carrot length.
If positive the correlation is in the same direction and if negative its opposite.

3.4 Coefficient of determination R^2

How well a statistical model predicts an outcome. A goodness of fit, how close a models variance shows the dependant variables variance.

Can use the Pearsons correlation coefficent or least squares estimate.

3.5 Alkaline Information Criteria

How well a model fits the data it was generated from.

The best model explains the greatest amount of variation with the fewest independent variables.

Method: Choose a number of models with different number of independent variables and determine which is the best.

$$AIC = 2(K) * 2ln(L) \tag{3.4}$$

K: Number of independent variables
L: Log likelihood estimate

3.6 Standard error

The error in the mean calculation:

$$SE = \frac{\sigma}{\sqrt{n}}$$

4 Distribution

4.1 General

$$X \sim N(\mu, \sigma^2) \tag{4.1}$$

The random variable "X", belongs to "~" a normal distribution "N" with a specific mean and variance.

4.1.1 What is a probability mass function?

PMF: For discrete variables, it can give a calculation of the probability of an exact value.

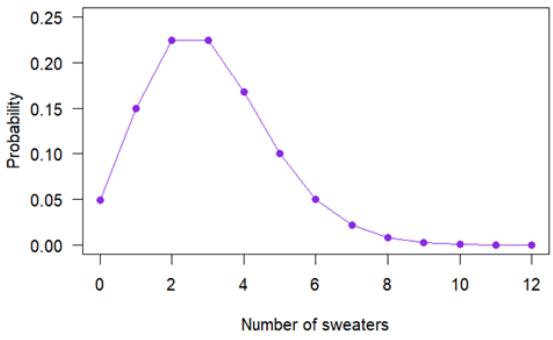

Figure 4.1: PMF

4.1.2 What is a probability density function ?

PDF(): For continous variables the probability of a single value is neglible and therefore assumed at zero, instead intervals probability is calculated. Instead for a given the probability density function is used, which measures the number of times a value is shown in a sample (its

Dist	**Description**
Binomial	Two states, the number of times one states shows in n trials
Bernoulli	Random variable is either one of two states
Discrete Uniform	probability of each state is equal
Poisson	Prob an event will happen k times in a given period of time or spaces

Table 4.1: PMF Dist

density).

Example: a carrot is 10cm, if the carrot shows as 10cm once in sample of 50, the PDF is 1/50. Can determine the probability by finding multiplying a interval by its PDF.

Dist	**Description**
Normal	Centered on the mean, bell shaped
Continuos Uniform	Equal intervals have equal probability
Log normal	Right skewed, normal when logged
Exponential	Higher prob for small values than large values.

Table 4.2: PDF Distributions

4.2 What is the bernoulli and binomial distribution?

Bernoulli distribution: The random variable can either be 0 or 1.

Binomial distribution: The random variable remains to have only two states, this shows the probability of measuring either state x number of times given n independent occurrences.

4.3 Poisson Distribution

4.4 What is the poisson distribution ?

- The random discrete variable is a count of the number of events occurring at random in regions of time and space. Ie radioactive particle emission or saplings in a sample of ground
  - All events are independent
  - No two events at the same time
  - Over a short period of time or on a small region the probability is the same

$$p_x = P(X = x) = e^{-\lambda} * \lambda^{x/x!} \tag{4.2}$$

Recurrence formula:

$$P(X = x) = \lambda / x * P(X = x - 1) \tag{4.3}$$

λ is the mean var and $\sqrt(\lambda)$ is the std
95% of values are between the mean ± 2 std
Independent Poisson random variables can be added to give another Poisson random variable

4.5 What is a normal distribution ?

Normal distribution : Data set centered evenly about a value, giving a bell curve.

4.5.1 How does std related to a normal dist?

STD	Percentage of values included (%)
σ	68
2σ	95
3σ	99.7

Table 4.3: std relation

4.5.2 What dist do multiple large mean samples show ?

They will show a normal dist, even if the variable itself doesn't show a normal dist.

4.6 Standard Normal Distribution

Also known as z dist is a normal dist with mean = 0 and std = 1. Any normal dist can be converted into a z dist, using the below formula to work out the z value (which calculates how many std vals a value x from the mean is).

$$z = \frac{x - \mu}{\sigma} \tag{4.4}$$

Great reference to use as all numbers have been calculated. Can allow:

- Comparisons of sample mean to population mean
- Compare different N dists (different mean and var)

4.7 Chi Squared Distribution

Not a reflection of real world distributions, but instead used for testing. Shaped by k (degrees of freedom) , made of squared z dist with different layers of multiple of std added.

$$\chi^2_k = (Z_1)^2 + (Z_2)^2 ... + (Z_{n\sigma})^2 \tag{4.5}$$

Hypothesis tests follow the chi squared dist under the null hypothesis. A commonly used tests is the Pearson chi squared test. There is also a non centered chi squared dist for any skewed data.

Goodness of fit tests measures how well a model fits a set of observations, there is also the chi squared Independence test.

4.8 What is a T Distribution?

If the the sample size is limited and below 30, then instead of a normal and T distribution is used.
A T dist has degrees of freedom (v) = n-1:

- A normal distribution has degrees of freedom v = ∞
- n being the sample size

If the variance is unknown and n is large, Z can be adjusted to use sˆ2 which is an unbiased estimate of the variance. This gives two random variables in the equation X and S:

$$T = (\bar{X} - \mu)/(S/\sqrt{n}) \tag{4.6}$$

c is the critical value depending on the distribution parameters and the confidence interval required:

$$(\bar{x} + c(s\sqrt{n}), \bar{x} - c(s\sqrt{n})) \tag{4.7}$$

5 Estimation

5.1 What is the confidence interval?

Mean of estimate ± variation in estimate

The interval is a range with a certainty value if the test was repeated in the same way the with a different sample, the result would be the same. Essentially its a method of minimizing the sample errors effect, however does not give a certainty on the "true value".

5.2 How is the confidence interval determined?

The interval is 1-α, where alpha is determined arbitrarily by the field of research.

5.3 What is the critical value?

It is how many stds you need to go from the mean, before you reach the confidence level. To calculate the critical value normally a transformation to a z dist is performed as the critical values are well known.

5.4 How do you calculate the critical value?

The critical value is the when the probability of the answer is between your interval. If converted to a z score, the value should always be less than z value at the chosen interval.

Use two tailed for a 2 dimensional comparison and one tail for a 1 dimensional comparison.

For a two tailed test looking for an interval of 95% the alpha is 0.025 (as double tailed) and with a normal dist the critical value is 1.96.

5.5 How do you calculate the confidence interval ?

If n observations are made from a N(μ, σ^2) dist, the random variable mean from each observation will have mean of μ and variance of $\frac{\sigma^2}{n}$.

To calculate the confidence interval of a sample mean, use the variable Z below (the critical value), where $\bar{X}$ is random variable corresponding to the sample mean:

$$Z = (\bar{X} - \mu)/(\sigma/\sqrt{n}) \tag{5.1}$$

This value should then be greater or less than the critical value, giving the result.
For a normal distribution with mean 0 and std of 1 N(0,1) the confidence interval (95%) is:

$$(\bar{x} + 1.96(\sigma/\sqrt{n}), \bar{x} - 1.96(\sigma/\sqrt{n})) \tag{5.2}$$

The above interval on an average of 95% of the time will include the mean.

5.6 How is s calculated ?

s is the **sample** std, not the population which is σ.

$$s^2 = 1/(n-1) * \sum (x - \bar{x})^2 \tag{5.3}$$

6 Statistical tests

6.1 What are stats tests used for?

- Determine predictor vs outcome relationship
- Estimate diff between two or more groups

The null hypothesis for any statistical test is that there is no relationship or difference between the two groups.

A number of assumptions regarding Independence, homogenity (similar variance) and normality (N()) are required to use **parametric tests**. If they are violated non parametric tests can be used.

6.2 How are test statistics used?

Test statistics measure the relationship between the variables and the null hypothesis.

Assuming the null hypothesis is true, the probability a more extreme (in a wrong way) test stat in the direction of the alternative is observed. If < alpha unlikely a extreme stat toward the alternative will be seen if > alpha (big) then likely.

If p small (< alpha) null rejected for alternative, essentially gives validation to the current alternative. If p big (> alpha) null cannot be rejected, unknown about the current alternative.

6.3 What is the p value?

P value is the probability that the observed test stat indicates the null hypothesis is true.

6.4 What are the types of values used in stats tests?

6.5 Parametric Tests

See table 6.2 on 13.

6.6 Nonparametric tests

See table 6.3 on 14.

Variable	**Description**
Continuos	quantitative real data
Discrete	quantitative integer data
Ordinal	Order based data ie rankings
Nominal	Names ie brand names
Binary	bits

Table 6.1: Variable types

7 Hypothesis testing

7.1 What are the two hypothesis statements?

2 hypothesis are given the null (H_o) and the alternative (H_1): - Null gives a specific parameter value - Alternative gives a range of values
Example of a parameter used is the population mean (μ)

7.2 How do you determine the confidence of a given null hypothesis ?

A normal distribution of N(μ, σ^2) can be related to N(0,1) by using $z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$.

A normal distribution if the variable is adjusted to be the sample mean $\bar{X}$ using the mean from the null hypothesis will become N($\mu_o, \sigma^2/n$).

This can then be used in the form:$P(\bar{X} > x) = P(z > (x - \mu)/\sigma)$;
This is then used to calculate the confidence percentage, from the tails of the distribution of a N(0,1).

Two tailed tests gives a much more complete analysis of the data set.

7.3 Example

A random sample of carrots are taken from a field, the defined length of carrot is 7.5 inch.

$$H_o : \mu = 7.5$$

$$H_A : \mu \neq 7.5$$

As the dist is of carrot sizes is likely normal, the sample size of N being 10 a t test is used.
If the significance level used an alpha of 0.05, the null hypothesis would be rejected if the t stat is less than -2.2616 or greater than 2.2616.

The t stat is 1.54 and therefore is not in the **critical region** and does not reject the null hypothesis.

The p value is the area under the graph above/below the critical value. The p value here is 0.158 which is greater than the alpha value showing the prob of it being outside the interval

7.4 What are some basic terms ?

Test statistic : Function of data used to determine between H_o andH_1
The critical region: Values that lead to rejection of H_o in favour of H_1 is the critical region.

Significance level: The probability H_o is rejected for H_1

7.5 What are the error types ?

Type 1 error: H_0 is rejected for H_1 however it was correct; This is mitigated by choosing a low significance level.
Type 2 error: H_o is accepted but incorrect.

7.6 What is the suggested test procedure ?

1. State the two hypothesis (Null and alternative)
2. Choose the appropriate test statistic and distribution
3. Choose significance level
4. Collect data
5. Analyze

To avoid bias the sig level should be chosen before any data is collected
If the dataset is approximately normal dist then use the standard N(0,1)

7.7 How does confidence level relate to significance level ?

If μ_0 is outside the range of α% confidence level, then the significance level is (100-α%)

7.8 Chi squared dist

Uses the standard v degrees of freedom.
Only used for non negative random variables (generally for freq measurements) to determine if two variables are dependant or independent. This includes if a variable is bias by comparing it to the expected non bias result.

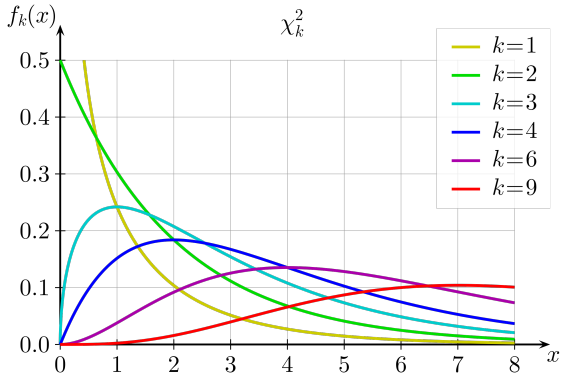

Figure 7.1: Chi square

As shown it is a skewed dist, as v increases the skew decreases.

7.9 How do you test for bias?

The difference between the expected (non bias result) and the observed result will indicate bias. Both size and relative size matter:

$$(O-E)*(O-E)/E = (O-E)^2/E \tag{7.1}$$

O: Observed, E: Expected

7.10 What value is used to determine goodness of fit between two models?

$$X^2 = \sum_{i=1}^{m} (O_i - E_i)^2 / E_i \tag{7.2}$$

Where m is the number of different outcomes for each model (columns).
Large value of X^2 suggest a lack of fit

7.11 How does X^2 relate to Chi squared ?

Chi squared dist approximately shows the probability distribution of X^2, if the freq values > 5 :

$$X^2 = \chi^2_{m-1} \tag{7.3}$$

7.12 What is a contingency table ?

A table with more than two variables being measured against (two+ rows)
The degree of freedom is : $v = (r-1)(c-1)$
r: rows, c: columns
If X^2 is within the chi squared 95% interval it should be accepted as independent.

7.13 What should you do with a 2x2 table ?

Can use the alternative:

$$X^2 = \sum (|O - E| - 0.5)^2 / E \tag{7.4}$$

Test	Predictor Var	Outcome Var	Ex
Simple Linear Regression	Continuos,1 predictor	Continuos, 1 outcome	How does moisture effect carrot size
Multiple Linear Regression	Continuos, 2+ predictor	Continuos, 1 outcome	How does moisture and temp effect carrot size
Logistic regression	Continuos, 1 predictor	Binary, 1 outcome	What is the effect of soil nutrient level on carrot death
Paired t test	Categorical, 1 predictor	Quantitative, set of population	What is the effect of two soil types on carrot mean length from one batch
Independent t test	Categorical, 1 predictor	Quantitative, sets from different population	Difference in mean carrot length from two farms
ANOVA	Categorical, 1+ predictor	Quantitative,1 outcome	Average carrot length with three different soil types
MANOVA	Categorical, 1+ predictor	Quantitative, 2+ outcomes	Average carrot length and width effected by soil type and seed type
Pearsons r	Continuos, 2+	Continuos, 1	How are carrot length and leaf length correlated

Table 6.2: Parametric tests

Test	Predictor Var	Outcome Var	Instead of
Spearmens r	Quantitative	Quantitative	Pearsons r
Sign test	Categorical	Quantitative	One sample t test
Chi square Independence	Categorical	Categorical	Pearsons r
Kruskal-Wallis H	Categorical, 3+	Quantitative	ANOVA
ANOISM	Categorical, 3+	Quantitative 2+	MANOVA
Wilcoxon Rank Sum	Categorical, 2	Quantitative, diff pops	Independent t test
Wilcoxon Signed-rank test	Categorical 2,	Quantitative, same pop	Paired t test

Table 6.3: Non parametric tests