Poisson – a discrete distribution

- Mean = λ
- Variance = λ

  a single parameter describes both variance and the mean

Negative binomial – a discrete distribution with a parameter *k* that provides an index of dispersion.

- Mean = μ
- Variance = $\mu + k\mu^2$

  the variance is greater than the mean

Log normal – a continuous distribution.

The logarithm of the values in this distribution are normally distributed.

Standard normal – a normal distribution with mean = 0 and variance =1

The distributions that we will be most concerned with are the normal and the standard normal.

Measures of dispersion

Our first major objective is to develop the concepts needed to understand hypothesis testing. We will primarily test hypotheses about means, but variances can also be tested. Testing means will require a measure of the dispersion or variability in the data set, so testing both means and variances requires knowledge of variance.

The following presents some measures of variation or variability among the elements (observations) of a data set

- Range – difference between the largest and smallest observation

  This is a rough estimator which does not use all of the information in the data set.

- Interquartile range – difference between the third and first quartile ($Q_3 - Q_1$)

  Recall that the first quartile (Q_1) is the value that has one quarter of the observations with lesser values and the third quartile has three quarters of the observations with lesser values. This may be a better measure of variability than the range in most situations because the range can be influenced by a single unusually large or unusually small value. However, this measure also does not use all of the information in the data set.

- Variance – the "average" squared deviation from the mean,

  The Population Variance is σ^2 (called "sigma squared")

  This is a parameter, and therefore a constant

  The variance is given by $\sigma^2 = \frac{\sum_{i=1}^{N}(Y_i - \mu)^2}{N}$ where *N* is the size of the population

  S^2 is the Sample Variance (called "S-squared").

  This is a statistic, and therefore a variable

The sample variance is given by $S^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$ where n is the size of the sample

NOTE that the divisor is n–1 rather than n. If n is used then the calculation is a biased estimator of σ^2, tending to be an under estimate.

Standard Deviation – a standard measure of the deviation of observations from the mean. It is calculated as the square root of the variance

$\sigma = \sqrt{\sigma^2}$ this is a parameter

$S = \sqrt{S^2}$ this is a statistic

Mean Absolute Deviation (MAD) – the "average deviation" from the mean, but using absolute values. This is another possible measure of dispersion. However the variance is the usual calculation as it has some advantages over the MAD.

Desirable properties of a measure of dispersion

A valid, useful measure of dispersion should:

- use all of the available information
- be independent of other parameters (and statistics) for large data sets
- be capable of being expressed in the same units as the variables
- be small when the spread among the points in the data set is small, and large when the spread is wider.

The Standard deviation meets these criteria.

A note on units

When we calculate the mean for a sample or population, the units on the mean are the same as for the original variable. If the original variable was measured in inches, the units of the mean will be inches

The variance also has units, but since the calculation involves the square of the original variable, the units on the variance are the original variable squared. If the original variable was measured in inches, the units of the variance would be inches squared

Since the standard deviation is the square root of the variance, the units on the standard deviation would again be the same as the original variable.

Degrees of freedom (d.f.)

In the calculation of a population variance the divisor is N, while in the calculation for a sample the divisor is n–1. This is because the calculated estimate of one parameter (σ^2) uses an estimate of another parameter (μ) in its calculation. For a sample, the estimate of the variance (S^2) employs a previously estimated statistic ($\bar{Y}$). Since we use an estimate of $\bar{Y}$ to calculate our estimate of S^2, the divisor is n–1, $S^2 = \frac{\Sigma\left(Y_i - \bar{Y}\right)^2}{n-1}$. This denominator is called its degrees of freedom. If we needed to estimate two parameters prior to estimating a parameter, the d.f. would be n–2.

Why? If we knew μ, as we do for a population, then we could get an independent deviation from each and every observation.

If we knew that $\mu = 5$, and we drew an observation at random and its value was 3, then the deviation would be –2. Each and every observation contributes a deviation since we know the value of μ.

But we cannot get an estimate of σ^2 from a single sample observation since that observation is also its own mean and the deviation is zero. If we drew a single sample observation, with a value of 3, and we did not know the value of μ, then we would estimate the value of $\bar{Y}$ from our sample. That estimated value would also be 3 and there would be no deviation.

In summary, with a known value of μ every observation can deviate independently from μ, and the sum of the deviations has no restrictions. However, deviations from $\bar{Y}$ always sum to ZERO, so only the first n–1 can assume "any" independent value. When we know the value of n–1 observations, the remaining observation is fixed by our knowledge of $\bar{Y}$.

Calculating the Variance

The variance is calculated as the sum of squared deviations divided by the degrees of freedom.

For a sample , $S^2 = \dfrac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$. This calculation requires going through the data once to estimate $\bar{Y}$ and a second time to estimate $(Y_i - \bar{Y})^2$.

The variance can, in many cases, be calculated more easily with the "calculator formula".

When we refer to "sum of squares", or SS, we will mean the "Corrected Sum of Squares", unless otherwise stated. When we need to refer to the uncorrected sums of squares they will be denoted as UCSS or USS.

Uncorrected sums of squares $\sum_{i=1}^{n}\left(Y_i^2\right)$

Corrected sums of squares (deviation formula) $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Corrected sums of squares (calculator formula) $\sum_{i=1}^{n}Y_i^2 - \dfrac{\left(\sum_{i=1}^{n}Y_i\right)^2}{n}$ or $\sum_{i=1}^{n}Y_i^2 - n\bar{Y}^2$

As noted, the deviation formula requires two passes through the data. However, since most calculators can simultaneously accumulate both the sum of Y_i, $\sum_{i=1}^{n}Y_i$, and the sum of Y_i squared, $\sum_{i=1}^{n}\left(Y_i^2\right)$, the calculation formula requires only a single pass through the data.

The "correction" made in corrected sums of squares is a correction for the mean. This is apparent in the deviation formula, but not a obvious in the calculator formula. The term $\left(\sum_{i=1}^{n} Y_i\right)^2 \Big/ n$ or $n\bar{Y}^2$ in the calculator formula is called the "correction factor", and corrects for the mean.

Finally, the sum of squares is divided by the degrees of freedom to get the variance. The value of the sum of squares should be the same regardless of the formula used.

An example of variance

Examine two samples;

Sample 1: 1, 2, 3 $\bar{Y} = 2$

Sample 2: 11, 12, 13 $\bar{Y} = 12$

Note that the deviations from the mean are the same in each case (–1, 0, 1) and the sum of squared deviations, $SS = (-1)^2 + (0)^2 + (1)^2 = 2$, is also the same for both samples using the deviation formula.

The corrected SS using the calculator formula are also the same

Sample 1 SS = 14 – 12 = 2

Sample 2 SS = 434 – 432 = 2

And the Variance for both samples is then SS / (n–1) = 2 / 2 = 1

So, two different looking sets of numbers have the same "scatter" and the same variance.

Coefficient of variation

CV is the standard deviation expressed as a percent of the mean, $CV = \left(S \big/ \bar{Y}\right) 100\%$

the CV is used to compare relative variation between different experiments or variables, independent of the mean. This calculation allows the comparison of different variables (variability on automobile weights to variability in hippopotamus weights) or variables on different scales (e.g. inches to kilograms).

Examples:

compare the variability of peoples weights to peoples heights.

compare variation in infants lengths to adult heights.

NUMERICAL Example: compare the relative variation in fork length of fish to the weights and scale lengths of the same fish. Data from 3 year old Flier Sunfish (*Centrarchus macropterus*).

	Length (mm)	Weight (g)	Scale Lt. (mm)
Mean	131.8	53.0	6.9
Std Dev	15.1	19.6	0.8

CV (length) = (15.1 / 131.8) × 100% = 11.5%

CV (weight) =(19.6 / 53.0) × 100% = 37.0%

CV (scale length) = (0.8 / 6.9) × 100% = 11.6%

From the results above we may conclude that the fish weights are relatively more variable than their length, and that the variability in body length and scale length are nearly the identical.

Note:

the CV has no units and can be highly variable and may well exceed 100%

From SAS example #1a

See SAS output Coefficient of Variation and other statistics discussed

Expected values and Bias

DEFINE

Unbiased Estimator: a statistic is said to be an unbiased estimator of a parameter if, with repeated sampling, the average of all of the sample statistics approaches the parameter. An estimator would be biased if on the average it approached a value that was a larger or smaller than the true target parameter.

Expected value: the mean value of a statistic from a large number of samples (the "long run" average). From our previous discussions, dividing by n–1 to calculate variance for a sample results in a value which is LARGER than if we divide by n. If dividing by n–1 is the correct approach (giving an unbiased estimate), it suggests that dividing by a larger number, n, causes a negative bias (a value which is, on the average, too SMALL). This is true.

It is true that the expected value of the sample mean is equal to the population parameter (i.e. $E\left(\overline{Y}\right) = \mu$, and it is also true that $E\left(S^2\right) = \sigma^2$, so these are unbiased estimators.

Note that for symmetric distributions, μ can also be estimated by the median, mode or midrange. However, the mean is an unbiased estimator for all distributions.

Expected Values are actually calculated as a sum (or integration for continuous variables) of the product of the observed values (Y_i) in the distribution and the probability ($p(Y_i)$) of occurrence of each value (e.g. $E(Z_i) = \Sigma[Z_i \times P(Z_i)]$). These have various uses, including the evaluation of bias.

For our purposes;

The expected value is the measure of the true central tendency for the probability distribution. If we took all possible samples, the mean would be the expected value, provided the estimator we used is unbiased.

For any statistic, if the expected value of the statistic is the same as the population value, the statistic is unbiased.

Summary of Dispersion

Dispersion is a measure of the variability among the elements of a population or sample

A number of estimates are available, including the Range, Interquartile range, Variance and Standard deviation. All are available from SAS PROC UNIVARIATE.

Units of the variable are squared on variances, but the same as the original variable for standard deviations.

Calculations on samples must consider degrees of freedom.

Both the sample means and sample variances (when divided by "n–1") are unbiased estimators of their target parameters, the population mean and population variance, respectively.

Constructing a Frequency Table

DIVIDE the population into a number of classes or groups based on the characteristics studied.

Categories are often quantitative, but not necessarily

DETERMINE the number of observations in each class (i.e. the frequency of occurrence of observations in each class).

CONSTRUCT the table with both classes and frequencies. The frequencies may also be relative (i.e. percentages) or cumulative.

Example

Construct a frequency table for a population of fish age groups.

N = 10

Y = age of fish in years: 8, 4, 4, 0, 1, 5, 6, 5, 3, 4

These values are placed into discrete age groups (0 to 8)

Frequency Table

Class value	Frequency (f.)	cumulative frequency (c.f.)
0	**1**	**1**
1	**1**	**2**
2	**0**	**2**
3	**1**	**3**
4	**3**	**6**
5	**2**	**8**
6	**1**	**9**
7	**0**	**9**
8	**1**	**10**
SUM	**10**	

Additional terms

Frequency Total: the total number of observations. The sum of the class frequencies.

Frequency (f): the number of observations in each class

Cumulative Frequency (c.f.): The sum of all class frequencies up to and including the class in question. Implies an order or rank, so this is usually done only with QUANTITATIVE VARIABLES

Relative Frequency (r.f.): the ratio of the class frequencies to the total frequency. These always sum to 1.0

r.f. * 100% gives the percentage frequency (sums to 100%)

Relative Cumulative Frequency (r.c.f.): the sum of the r.f. up to and including the class in question (for QUANTITATIVE VARIABLES).

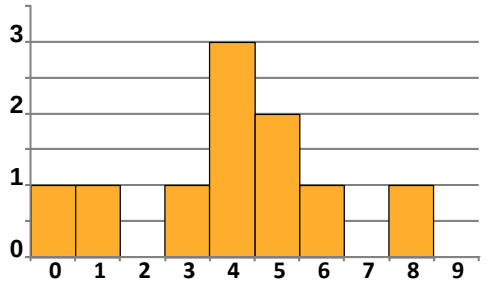
Frequency Table

Class value	frequency	Cumulative frequency	Relative frequency (r.f.)	relative cumulative frequency (r.c.f)
0	1	1	0.1	0.1
1	1	2	0.1	0.2
2	0	2	0.0	0.2
3	1	3	0.1	0.3
4	3	6	0.3	0.6
5	2	8	0.2	0.8
6	1	9	0.1	0.9
7	0	9	0.0	0.9
8	1	10	0.1	1.0
SUM	10		1.0	

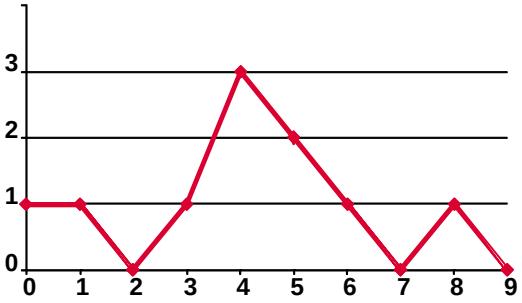
Graphic displays of frequencies

HISTOGRAM or bar-chart - representation of a frequency table

The area under each bar is proportional to the relative frequency (r.f.) of the class.



FREQUENCY POLYGON a variation of a histogram type plot in which the midpoints of each class relative frequency is connected with a straight line.



Characteristics of histograms

When done with relative frequencies, the total area of a graph of relative frequencies is 1.0

Any subsection of a graph of relative frequencies will have an area such that, $0 \leq$ subsection area ≤ 1

SAS example (#1b) from Freund & Wilson (1997) Table 1.1, **see SAS output for results**

Things to note – Options

```
dm'log;clear;output;clear';
OPTIONS LS=99 PS=512 nocenter nodate nonumber;
ODS HTML body='C:\Example01.html' ;
TITLE1 'Introductory SAS example 1';
```

– the DATA step
– the raw DATA (note ending semicolon)
– the Procedures

```
PROC MEANS
PROC SORT; BY QUALITY;
PROC MEANS; BY QUALITY;
PROC FREQ
PROC CHART; VBAR QUALITY;
PROC CHART; HBAR QUALITY;
proc gchart; pie QUALITY;
proc gchart; star QUALITY;
proc gchart; donut QUALITY;
```

Summary

Frequencies are a common and useful technique for descriptive statistics with many possible presentations.

We would usually do the calculations in SAS

The distributions that we will use for hypothesis testing will be in the form of frequency distributions

Linear Models

The simplest form of the linear additive model

$$Y_i = \mu + \varepsilon_i \quad \text{for i = 1, 2, 3,...,N}$$

This is a population version of the model, so the term μ is a constant, it is the population mean

The sample version would use $\overline{Y}$, which is a statistic and a variable.

ε_i represents the deviations of the observations from the mean. It has a mean of zero since deviations sum to zero.

e_i would be used to represent sample deviations,

and, of course, the population size, N, would be changed to the sample size, n.

This is a mathematical representation of a population or sample. All of the analyses discussed in the Statistical Methods courses have a linear model. The models get more complex as the analysis gets more advanced.

Multiplicative models and multiplicative errors exist, but are not covered in basic statistical methods. Note that the error term in this model is additive.

Other models we will discuss this semester include:

$Y_{ij} = \mu_i + \varepsilon_{ij}$ ..... for the two sample t-tests:

$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ ..... another form of the t-test also used for ANOVA

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ ..... Simple Linear Regression

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$ ..... Multiple Linear Regression

Coding and Transformations

Objective – Hypothesis testing Background

Many applications in statistics require modifying an existing distribution to an alternative form of the distribution. Hypothesis testing, in particular, requires taking an observed distribution and transforming to a recognized statistical distribution with known properties. This modification involves a transformation.

Theorems

If a constant "a" is added to each observation then, the mean of the data set will increase by "a" units the variance and standard deviation will remain unchanged

Example: Population of size N = 4

$Y_i = 2, 4, 6, 8$

$$\mu = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{20}{4} = 5$$

$$\sigma_Y^2 = \frac{\sum_{i=1}^{N} (Y_i - \mu)^2}{N} = \sum_{i=1}^{N} Y_i^2 - \frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N} \Bigg/ N = \frac{(120 - 100)}{4} = 5$$

$$\sigma_Y = 2.24$$

Now add 10 to each observation, the population size is still N = 4

$Y_i = 12, 14, 16, 18$

$$\mu = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{60}{4} = 15$$

$$\sigma_Y^2 = \sum_{i=1}^{N} Y_i^2 - \frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N} \Bigg/ N = \frac{(920 - 900)}{4} = 5$$

$$\sigma_Y = 2.24$$