

## Probability and Statistics Review

Population vs. sample:  $N$  vs.  $n$

Experimental vs. observational studies: in experiments, we manipulate the results whereas in observational studies we simply measure what is already there.

Variable of interest/ dependent variable/ response variable/ outcome:  $y$

Auxiliary variables/ explanatory variables/ predictor variables/ independent variables/ covariates:  $x$

Observations: Measure  $y$ 's and  $x$ 's for a census (all  $N$ ) or on a sample ( $n$  out of the  $N$ )

$x$  and  $y$  can be: 1) continuous (ratio or interval scale); or 2) discrete (nominal or ordinal scale)

Descriptive Statistics: summarize the sample data as means, variances, ranges, histograms, etc.

Inferential Statistics: use the sample statistics to estimate the parameters of the population

## Parameters for populations:

1. Mean --  $\mu$  e.g. for  $N=4$  and  $y_1=5; y_2=6; y_3=7, y_4=6$   
 $\mu=6$

2. Range: Maximum value – minimum value

3. Standard Deviation  $\sigma$  and Variance  $\sigma^2$

$$\sigma^2 = \frac{\sum_{i=1}^N (y_i - \mu)^2}{N}$$
$$\sigma = \sqrt{\sigma^2}$$

4. Covariance between  $x$  and  $y$ :  $\sigma_{xy}$

$$\sigma_{xy} = \left( \frac{\sum_{i=1}^N (y_i - \mu_y)(x_i - \mu_x)}{N} \right)$$

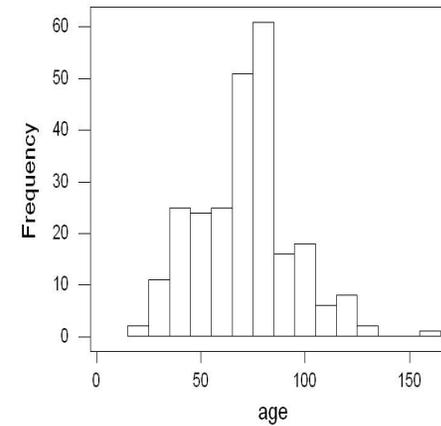
5. Correlation (Pearson's) between two variables,  $y$  and  $x$ :  $\rho$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \times \sigma_y^2}}$$

Ranges from -1 to +1; with strong negative correlations near to -1 and strong positive correlations near to +1.

6. Distribution for y -- frequency of each value of y or x (may be divided into classes)
7. Probability Distribution of y or x – probability associated with each y value
8. Mode -- most common value of y or x
9. Median -- y-value or x-value which divides the distribution (50% of N observations are above and 50% are below)

Example: 250 *Populus* trees of Alberta



**Descriptive Statistics: age**

N=250 trees      Mean = 71 years      Median = 73 years

25% percentile = 55      75% percentile = 82

Minimum = 24      Maximum = 160

Variance = 514.7      Standard Deviation = 22.69

1. Compare mean versus median
  2. Normal distribution?
- Pearson correlation of age and dbh = 0.573 for the population of N=250 trees

Statistics from the Sample:

1. Mean --  $\bar{y}$  e.g. for  $n=3$  and  $y_1=5; y_2=6; y_3=7$ ,  
 $\bar{y}=6$

2. Range: Maximum value – minimum value

3. Standard Deviation  $s$  and Variance  $s^2$

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{(n-1)}$$
$$s = \sqrt{s^2}$$

4. Standard Deviation of the sample means (also called the Standard Error, short for Standard Error of the Mean) and its square called the variance of the sample means are estimated by:

$$s_{\bar{y}}^2 = s^2/n \quad \text{and} \quad s_{\bar{y}} = \sqrt{s^2/n}$$

5. Coefficient of variation (CV): The standard deviation from the sample, divided by the sample mean. May be multiplied by 100 to get CV in percent.

6. Covariance between  $x$  and  $y$ :  $s_{xy}$

$$s_{xy} = \left( \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) \right) / (n-1)$$

7. Correlation (Pearson's) between two variables,  $y$  and  $x$ :  $r$

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 \times s_y^2}}$$

Ranges from -1 to +1; with strong negative correlations near to -1 and strong positive correlations near to +1.

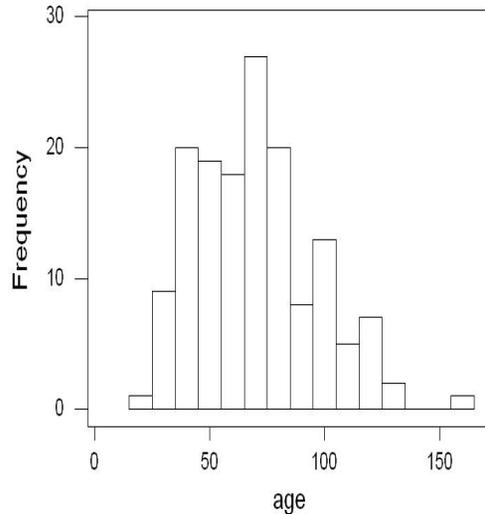
8. Distribution for  $y$  -- frequency of each value of  $y$  or  $x$  (may be divided into classes)

9. Estimated Probability Distribution of  $y$  or  $x$  – probability associated with each  $y$  value based on the  $n$  observations

10. Mode -- most common value of  $y$  or  $x$

11. Median --  $y$ -value or  $x$ -value which divides the estimated probability distribution (50% of  $N$  observations are above and 50% are below)

Example: n=150



n=150 trees      Mean = 69 years    Median = 68 years  
25% percentile = 48    75% percentile = 81  
Minimum = 24      Maximum = 160  
Variance = 699.98    Standard Deviation = 25.69 years  
Standard error of the mean = 2.12 years  
Pearson correlation of age and dbh = 0.66 with a p-value of  
0.000 for the sample of n=150 trees from a population of  
250 trees

**Good estimate of population values?**

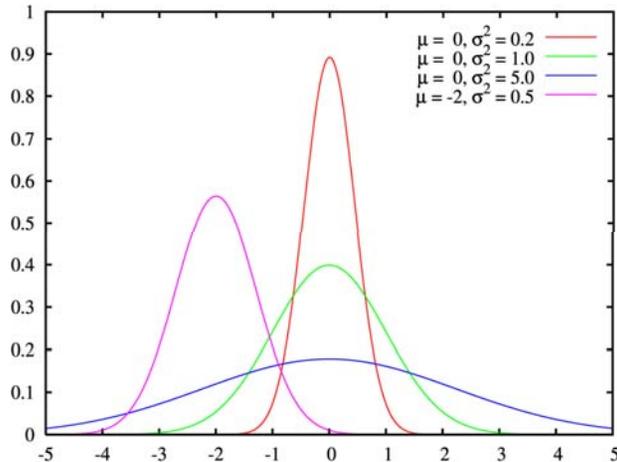
Sample Statistics to Estimate Population Parameters:

If simple random sampling (every observation has the same chance of being selected) is used to select n from N, then:

- Sample estimates are **unbiased estimates** of their counterparts (e.g., sample mean estimates the population mean), meaning that over all possible samples the sample statistics, averaged, would equal the population statistic.
- A particular sample value (e.g., sample mean) is called a “**point estimate**” -- do not necessarily equal the population parameter for a given sample.
- Can calculate an interval where the true population parameter is likely to be, with a certain probability. This is a **Confidence Interval**, and can be obtained for any population parameter, IF the distribution of the sample statistic is known.

## Common continuous distributions:

Normal:



- Symmetric distribution around  $\mu$
- Defined by  $\mu$  and  $\sigma^2$ . If a variable has a normal distribution, and we know these parameters, then we know the probability of getting an interval for any particular values of the variable.
- Probability tables are for  $\mu=0$  and  $\sigma^2=1$ , and are often called z-tables.
- Examples:  $P(-1 < z < +1) = 0.68$ ;  $P(-1.96 < z < 1.96) = 0.95$ .  
Notation example: For  $\alpha=0.05$ ,

$$z_{\alpha/2} = z_{0.025} = -1.96$$

- z-scores: scale the values for y by subtracting the mean, and dividing by the standard deviation.

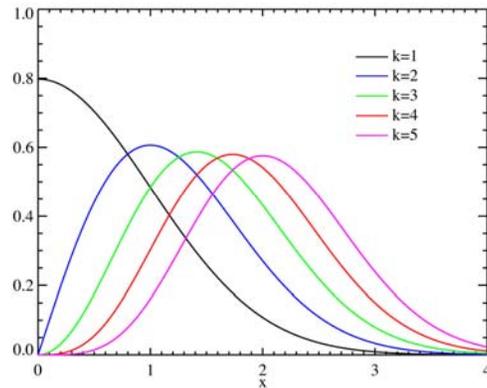
$$z_i = \frac{y_i - \mu}{\sigma}$$

E.g., mean=20, standard deviation of 2, y=10, z=-5.0

t-distribution:

- Symmetric distribution
- Table values have the center at 0. The spread varies with the *degrees of freedom*. As the sample size increases, the df increases, the spread decreases, and will approach the normal distribution.
- Used for a normally distributed variable whenever the variance of that variable is not known.
- Notation examples:  $t_{n-1, 1-\alpha/2}$  where n-1 is the degrees of freedom, in this case, and we are looking for the  $1-\alpha/2$  percentile. For example, for n=5 and  $\alpha=0.05$ , we are looking for t with 4 degrees of freedom and the 0.975 percentile (will be a value around 2).

$\chi^2$  distribution:



- Starts at zero, and is not symmetric
- Is the square of a normally distributed variable e.g. sample variances have a  $\chi^2$  distribution if the variable is normally distributed
- Need the degrees of freedom and the percentile as with the t-distribution

F-distribution:

- Is the ratio of 2 variables that each have a  $\chi^2$  distribution eg. The ratio of 2 sample variances for variables that are each normally distributed.
- Need the percentile, and two degrees of freedom (one for the numerator and one for the denominator)

Central Limit Theorem: As  $n$  increases, the distribution of sample means will approach a normal distribution, even if the distribution is something else (e.g. could be non-symmetric)

Tables in the Textbooks:

Some tables give the values for probability distribution for the degrees of freedom, and for the percentile. Others give this for the degrees of freedom and for the alpha level (or sometimes  $\alpha/2$ ). Must be careful in reading probability tables.

Confidence Intervals for a single mean:

➤ Collect data and get point estimates:

- The sample mean,  $\bar{y}$  to estimate of the population mean  $\mu$  ---- Will be unbiased
- The sample variance,  $s^2$  to estimate of the population variance  $\sigma^2$  ---- Will be unbiased

➤ Can calculate interval estimates of each point estimate

e.g. 95% confidence interval for the true mean

- If the  $y$ 's are normally distributed OR
- The sample size is large enough that the Central Limit Theorem holds --  $\bar{y}$  will be normally distributed

$n$  items measured out of  $N$  possible items  
(sometimes  $N$  is infinite)

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} \quad \text{where} \quad \sum_{i=1}^n y_i \quad (\text{sum over all } n \text{ items})$$

$$\sum_{i=1}^n y_i^2 \quad (\text{square each value and then add them})$$

$$s_y^2 = \frac{\sum y_i^2 - (\sum y_i)^2 / n}{n - 1}$$

$$s_{\bar{y}}^2 = \frac{s_y^2}{n} \left( \frac{N - n}{N} \right) \text{ without replacement;}$$

$$s_{\bar{y}}^2 = \frac{s_y^2}{n} \text{ with replacement or}$$

when  $N$  is very large

$$\text{Coefficient of Variation} = CV = \frac{s_y}{\bar{y}} \times 100$$

95% Confidence Intervals for the true mean of the population :

$$\bar{y} + / - t_{n-1, 1-\alpha/2} \times s_{\bar{y}}$$

Examples:

n is: 4

Plot	volume	ba/ha	ave. dbh
1	200	34	50
2	150	20	40
3	300	40	55
4	0	0	0

mean:	162.50	23.50	36.25
variance:	15625.00	315.67	622.92
std.dev.:	125.00	17.77	24.96
std.dev. of mean:	62.50	8.88	12.48
t should be:	3.182		
Actual 95% CI (+/-):	198.88	28.27	39.71

<b>NOTE:</b>			
<b>EXCEL:</b>	<b>122.50</b>	<b>17.41</b>	<b>24.46</b>
<b>95%(+/-)</b>			
<b>t:</b>	<b>1.96</b>	<b>1.96</b>	<b>1.96</b>
	<b>not</b>		
	<b>correct!!!</b>		

Hypothesis Tests:

- Can hypothesize what the true value of any population parameter might be, and state this as null hypothesis (H0: )
- We also state an alternate hypothesis (H1: or Ha: ) that it is a) not equal to this value; b) greater than this value; or c) less than this value
- Collect sample data to test this hypothesis
- From the sample data, we calculate a sample statistic as a point estimate of this population parameter and an estimated variance of the sample statistic.
- We calculate a “test-statistic” using the sample estimates
- Under H0, this test-statistic will follow a known distribution.
- If the test-statistic is very unusual, compared to the tabular values for the known distribution (e.g., it may be VERY large), then the H0 is very unlikely and we conclude H1:

Example: Comparing two means:

We believe that the average weight of male ravens (birds) differs from female ravens

$$H_0: \mu_1 = \mu_2 \text{ or } \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 \neq \mu_2 \text{ or } \mu_1 - \mu_2 \neq 0$$

A sample of 20 birds is taken and each bird is weighed and released. 12 birds were males with an average weight of 1.2 kg and a standard deviation of 0.02 kg. 8 birds were females with an average weight of 0.8 and a standard deviation of 0.01 kg.

Means?

Sample Variances?

Test statistic:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{s_{\bar{y}_1 - \bar{y}_2}} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}}$$

t =

Under H0: this will follow a t-distribution with df = (n1+n2-2).

Find t-value from tables and compare, or use the p-value:

Conclude?

Errors for Hypothesis Tests and Power

	H0 True	H0 False
Accept	1- $\alpha$	$\beta$ (Type II error)
Reject	$\alpha$ (Type I error)	1- $\beta$

Type I Error: Reject H0 when it was true. Probability of this happening is  $\alpha$

Type II Error: Accept H0 when it is false. Probability of this happening is  $\beta$

Power of the test: Reject H0 when it is false. Probability of this is 1- $\beta$

What increases power?

- Increase sample sizes, resulting in lower standard errors
- A larger difference between mean for H0 and for H1
- Increase alpha. Will decrease beta.