

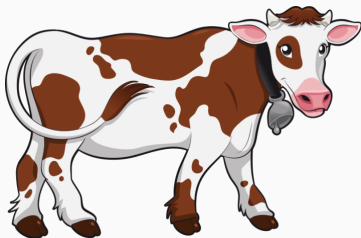
Module 1 - Intro and review

STAT 401

Section 1: Overview

What are the big ideas of statistical linear modeling?

- Statistics is the science of learning from empirical observation
- Linear models approximate reality to provide estimation of fixed quantities and prediction of random ones
- Association between variables and its uncertainty can be quantified



- Name: Buttercup
- Age: 3 yr.
- Weight: 1,015 lb.
- Price: \$1,000







- Name: Thistle
- Age: 2 yr.
- Weight: 1,025 lb.
- Price: \$1,175



- Name: Rose
- Age: 8 yr.
- Weight: 1,250 lb.
- Price: \$775



- Name: Petunia
- Age: 4 yr.
- Weight: 1,050 lb.
- Price: ?

Profile photo	Name	Age	Weight	Price
	Buttercup	3	1,015	1,000
	Thistle	2	1,025	1,175
	Rose	8	1,250	775
	Petunia	4	1,050	?

Section 2: Statistics

We start with some important terms:

Population: a group of individuals which are of interest

Sample: a subset of the population selected for study

Statistic: a numerical summary of a sample

Parameter: a numerical summary of a population

For example: an ornithologist studies all glaucous-winged gulls on earth. In order to estimate the relationship between wingspan and wing chord for them, he measures wing chord and wing span on four individuals and calculates the correlation coefficient.

Population: all glaucous-winged gulls on earth

Sample: the four gulls he studied

Statistic: sample correlation coefficient

Parameter: population correlation coefficient

For example: a dietician develops a diet for building muscle mass in athletes. She recruits 50 athletes and randomly assigns 30 to follow the diet and 20 to eat a control diet. After 8 weeks, she measures each person's change in muscle mass and calculates the average change for each diet.

Population: all athletes

Sample: the 50 recruited athletes

Statistic: sample average change in muscle mass

Parameter: population average change in muscle mass

For example: a trapper builds a snare for wolves in Yukon flats. After 6 attempts, he has caught one wolf.

Population: all possible attempts with the snare

Sample: the six attempts made

Statistic: the proportion of attempts that were
successful

Parameter: the probability of success over the long run

These statistics are some of the most commonly used:

- Sample mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Sample variance:

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Sample standard deviation:

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{s_x^2}$$

- Sample covariance:

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

- Sample correlation:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x s_y}$$

For the glaucous-winged gulls, the data are

wingspan (x)	wing chord (y)
130	42
126	40
127	41
129	41

$$\bar{x} = 128$$

$$\bar{y} = 41$$

$$s_x^2 = 10/3$$

$$s_y^2 = 2/3$$

$$s_x = \sqrt{10/3}$$

$$s_y = \sqrt{2/3}$$

$$s_{xy} = 4/3$$

$$r_{xy} \doteq 0.9$$

Remember that:

- $-\infty < s_{xy} < \infty$

- $s_x^2 \geq 0$

- $-1 \leq r_{xy} \leq 1$

Section 3: Random variables

Random variable: A function that assigns a number to each outcome of

Discrete random variable: A R.V. that only takes countably many values

Continuous random variable: A R.V. that takes uncountably many values

For example, X is a random variable which is 1 half the time, 2 one-third of the time, and 15 one-sixth of the time. X is discrete.

$$P(X = 1) = 1/2$$

$$P(X = 2) = 1/3$$

$$P(X = 15) = 1/6$$

$$P(X = x) = f(x), \text{ for some function } f$$

For example, Y is a random variable which is in the interval $[0, 1]$ two-thirds of the time and in $[1, 7]$ one-third of the time.

Y is continuous.

$$P(0 \leq Y \leq 1) = 2/3$$

$$P(1 \leq Y \leq 7) = 1/3$$

$$P(Y = 4) = P(Y = y) = 0$$

$$P(0.5 \leq Y \leq 2) = \text{uncertain}$$

Expectation: $E(X)$ is the mean of X and is a parameter

Variance: $V(X) = E[X - E(X)]^2 = E(X^2) - E(X)^2$

and is a parameter

Covariance: $Cov(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$

and is a parameter

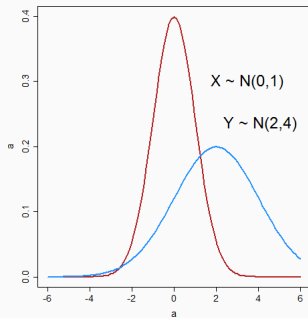
Correlation: $\rho_{XY} = Cor(X, Y) = \frac{Cov(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$

and is a parameter. If X and Y are independent,

$$\rho_{XY} = Cov(X, Y) = 0$$

Distribution: a function that assigns probabilities to the various values of a random variable

Normal distribution: a distribution for continuous R.V.s which resembles a bell. It is defined uniquely by two parameters, μ which is its mean, and σ^2 , which is its variance. If X is normally distributed with mean μ and variance σ^2 , this is often abbreviated $X \sim N(\mu, \sigma^2)$.



$$E(X) = 0$$

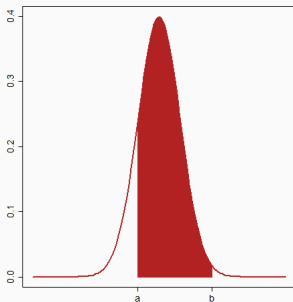
$$V(X) = 1$$

$$E(Y) = 2$$

$$V(Y) = 4$$

$$\text{Cov}(X, Y) = \text{uncertain}$$

Since X is continuous, $P(X = x) = 0$ for all x but we can meaningfully talk about $P(a \leq X \leq b)$ for constants a and b . This corresponds to the area under the normal density curve between a and b .



If $X \sim N(0, 1)$, then by the *empirical rule*, we know that

$$P(0 \leq X \leq \infty) = 0.5$$

$$P(-1 \leq X \leq 1) = 0.68$$

$$P(0 \leq X \leq 1) = 0.34$$

$$P(0 \leq X \leq 2) = 0.475$$

$$P(-0.25 \leq X \leq 1.6) = \text{uncertain}$$

In general, probabilities for normally distributed R.V.s must be looked up in a table or with software. It used to be necessary to standardize them first:

$$Z = \frac{X - \mu}{\sigma}$$

so that $Z \sim N(0, 1)$.

For example, if $Y \sim N(2, 4)$, then

$$P(Y \geq 0) = P\left(\frac{Y - 2}{2} \geq \frac{0 - 2}{2}\right) = P(Z \geq -1)$$

$$P(-2 \leq Y \leq 3) = P\left(\frac{-2 - 2}{2} \leq \frac{Y - 2}{2} \leq \frac{3 - 2}{2}\right)$$

$$= P(-2 \leq Z \leq 1/2)$$

The following properties are true of expectations and variances. For random variables X, Y and constants a, b, c :

$$E(a) = a$$

$$V(a) = 0$$

$$E(aX) = aE(X)$$

$$V(aX) = a^2V(X)$$

$$\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$$

Also,

$$E(X + Y) = E(X) + E(Y)$$

$$V(X + Y) = V(X) + V(Y) + 2\text{Cov}(X, Y)$$

$$E(X - Y) = E(X) - E(Y)$$

$$V(X - Y) = V(X) + V(Y) - 2\text{Cov}(X, Y)$$

Let $E(X) = 2$ and $E(Y) = -2$. Let $V(X) = 3$ and $V(Y) = 100$.

Let $\text{Cov}(X, Y) = 1$.

$$E(2X) = 2E(X) = 4$$

$$E(X + Y) = E(X) + E(Y) = 0$$

$$E(5X - 2Y) = 5E(X) - 2E(Y) = 14$$

$$V(2X) = 2^2V(X) = 12$$

$$V(X + Y) = V(X) + V(Y) + 2\text{Cov}(X, Y) = 105$$

$$V(5X - 2Y) = 5^2V(X) + 2^2V(Y) - 2 * 5 * 2\text{Cov}(X, Y) = 455$$

These properties partially help us prove the following facts about linear combinations of normal R.V.s. If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ and if X and Y are independent:

$$W = X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

$$V = X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

$$U = aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$$

for constants a, b .

Let $X \sim N(2, 3)$ and $Y \sim N(-2, 100)$ and let X and Y be independent. Then

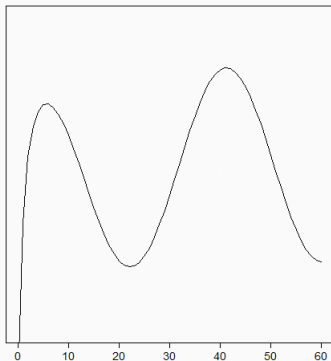
$$W = X + Y \sim N(0, 103)$$

Section 4: Sampling distributions

Sampling distribution of a statistic X : the distribution of X over repeated draws of a random sample from the population

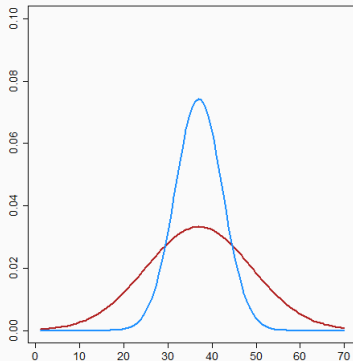
If \bar{X} is the mean age of a sample of people of size 5, then over repeated draws of samples of size 5, \bar{X} will have some sampling distribution:

Sample #	\bar{x}
1	42.1
2	31.0
3	19.7
4	57.8
\vdots	



Some populations are normally distributed with a mean μ and variance σ^2 . In such cases, the sampling distribution of \bar{X} is exactly $N(\mu, \sigma^2/n)$, or equivalently, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$.

If the mean population age is 37 and the variance is 144 and if ages are normally distributed, then the mean \bar{X} of a sample of size 5 has a sampling distribution of $N(37, 144/5)$ or $\frac{\bar{X}-37}{12/\sqrt{5}} \sim N(0, 1)$.



Central Limit Theorem (C.L.T.): Without (many)

conditions on the distribution of the values of individuals in the population, for a large sample size n , the sampling distribution of \bar{X} is approximately $N(\mu, \sigma^2/n)$, or $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$.

No matter how ages are distributed with mean 37 and variance 144 in the population, the sampling distribution of \bar{X} of sample size 50 is approximately $N(37, 144/5)$, or $\frac{\bar{X}-37}{12/\sqrt{5}} \sim N(0, 1)$.

Standard error: The standard error of a statistic is the standard deviation of its sampling distribution. Sometimes when we say “standard error” we mean “estimated standard error” because the C.L.T. is implicitly being used and/or reliance on s^2 instead of σ^2 .

If $\bar{X} \sim N(37, 144/50)$, then the standard error of \bar{X} or $S.E.(\bar{X})$ is $12/\sqrt{50}$.

Standard errors are a simple measurement of the precision of a statistic or estimator. Low standard errors are almost always desirable.

When we estimate σ^2 with s^2 , it is not, in general, the case that

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0, 1)$$

For normally-distributed populations, however, it is the case that

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

where $n - 1$ is the degrees of freedom of the distribution.

If the population of ages has mean 37, variance 144, and $n = 30$, then

$$\begin{aligned}P(16 \leq \bar{X} \leq 36) &= P\left(\frac{16 - 37}{12/\sqrt{30}} \leq \frac{\bar{X} - 37}{12/\sqrt{30}} \leq \frac{36 - 37}{12/\sqrt{30}}\right) \\&= P(-12.4 \leq Z \leq -0.6) = 0.274\end{aligned}$$

If the population of ages is normally distributed with mean 37 and $n = 50$, then

$$P(-12.4 \leq T \leq -0.6) = 0.277$$

Section 5: Confidence intervals

$(1 - \alpha) * 100\%$ **confidence interval:** A set of values which contain the true value of a parameter
 $(1 - \alpha) * 100\%$ of the time

$(1 - \alpha) * 100\%$ **prediction interval:** A set of values which contain the true value of a new realization of a random variable $(1 - \alpha) * 100\%$ of the time

If the parameter θ is to be estimated, the general form of a confidence interval is:

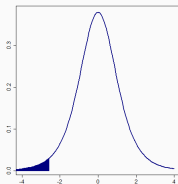
$$\text{estimate of } \theta \pm \text{multiplier} * \text{SE}(\text{estimate of } \theta)$$

The multiplier depends on the distribution of the standardized estimator of θ .

The general form of many C.I.s is:

$$\bar{x} \pm t_{\alpha/2} * \frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ is the $\alpha/2$ th tail of the t_{n-1} distribution.



Let the mean age of a random sample of size $n = 50$ be 35.1 and the sample variance be 144. What is a 95% confidence interval for the population mean?

$$35.1 \pm (-2) * \frac{12}{\sqrt{50}} = (31.7, 38.5)$$

In general, the following hold:

- As $(1 - \alpha) * 100\%$ (confidence level) increases, margin of error increases
- As $(1 - \alpha) * 100\%$ (confidence level) decreases, margin of error decreases
- As n increases, margin of error decreases
- As n decreases, margin of error decreases

If the sample size had been 30 instead of 50, the interval becomes

$$35.1 \pm (-2) \frac{12}{\sqrt{30}} = (30.7, 39.5)$$

If we had wanted 98% confidence, the interval becomes

$$35.1 \pm (-2.4) \frac{12}{\sqrt{50}} = (31.0, 39.2)$$

A common interpretation of a confidence interval states:

*We are $(1 - \alpha) * 100\%$ confident that the [parameter] is between [lower bound] and [upper bound]*

A more formal interpretation states:

*About $(1 - \alpha) * 100\%$ of such intervals calculated, over repeated sampling, would contain the true value of [parameter]*

We are 95% confident that the mean age of the population is between 31.7 and 38.5.

Section 6: Hypothesis tests

α -level hypothesis test: a formal method for defeating “null” (or default) claims about a parameter of interest so that it falsely defeats true null claims only $(100 * \alpha)\%$ of the time

Statistical hypotheses pit a null claim about a parameter against all other claims. In many cases:

- H_0 : the parameter equals some value
- H_A : the parameter does not equal that value

We can test scientific hypotheses by converting them into statistical hypotheses about some parameter. Once data is collected, we can calculate the probability of observing such data assuming that H_0 is true. Since we actually did observe such data, we might expect that probability to be high. If it turns out to be low, we conclude that something is wrong: namely, the assumption that H_0 is true.

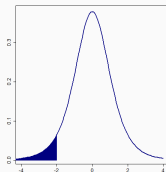
To test a null hypothesis, we follow five steps:

1. Formulate statistical hypotheses
2. Calculate a test statistic
3. Calculate a p -value
4. Decide whether to reject H_0 or not based on whether
$$p < \alpha$$
5. State your conclusion in the problem context

For testing a population mean μ , the steps will be:

1. $H_0 : \mu = \mu_0$ vs. $H_A : \mu \overset{\leq}{\neq} \mu_0$
2. $t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$ when H_0 is true

3. $p\text{-value} = P(T \text{ is more extreme than } t \text{ if } H_0 \text{ is true})$

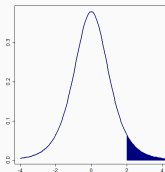


$$H_A : \mu < \mu_0$$

$p\text{-value} =$

$P(T < t \text{ if}$

$H_0 \text{ true})$

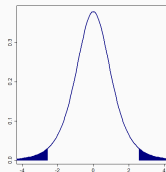


$$H_A : \mu > \mu_0$$

$p\text{-value} =$

$P(T > t \text{ if}$

$H_0 \text{ true})$



$$H_A : \mu \neq \mu_0$$

$p\text{-value} =$

$P(T < -|t| \text{ or}$

$T > |t| \text{ if } H_0 \text{ true})$

4. If $p\text{-value} \leq \alpha$, reject H_0

If $p\text{-value} \geq \alpha$, fail to reject H_0

5. State:

*There [is/is not] sufficient evidence at the $[\alpha*100\%]$ level to conclude that the mean of [variable] is [less than/greater than/not equal to] [value of μ_0]*

Suppose we wish to show that the population mean age is less than 37. For $n = 50$ we get $\bar{x} = 35.1$ and $s^2 = 144$.

1. $H_0 : \mu = 37$ vs. $H_A : \mu < 37$
2. $t = \frac{35.1-37}{12/\sqrt{50}} = -1.1$ and $T \sim t_{49}$ under H_0
3. $p\text{-value} = P(T < -1.1 \text{ if } \mu = 37) = 0.13$

4. If $\alpha = 0.05$ then $p\text{-value} = 0.13 > 0.05 = \alpha$; fail to reject H_0
5. There is not sufficient evidence at the 5% level to conclude that the population mean age is less than 37.

The more formal way to interpret a p -value to state:

If H_0 is true, there is a [p-value] probability that we would collect data that produces a test statistic as extreme or more extreme than the one we actually obtained.

If $\mu = 37$ in reality, there is a 0.13 probability of collecting a sample of size 50 from the population with a t statistic of -1.1 or lower (ie a sample mean age of 35.1 or lower)

Type I error: rejecting H_0 when it is true (due to sampling variability)

Type II error: failing to reject H_0 when it is false (due to sampling variability)

- $P(\text{Type I error}) = \alpha = \text{level of test}$
- $P(\text{Type II error}) = \beta = 1 - \text{power of test}$

In general, the following hold:

- As n increases and α is fixed, β decreases and power increases
- As n decreases and α is fixed, β increases and power decreases
- As α increases, β decreases and power increases
- As α decreases, β increases and power decreases