STATISTICS

PARAMETER ESTIMATION

CONFIDENCE INTERVAL

- Then the $100(1-\alpha)\%$ CI for μ is:

MEAN W 1 SAMPLE

KNOWN POP VAR
· Population variance σ^2 $\quad \bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$

UNKNOWN POP VAR
· Sample variance s^2 $\quad \bar{x} \pm t_{\nu,\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$

Have?	Sample
size	n
mean	$\bar{x}$

where t has $\nu = (n-1)$ degrees of freedom.

Then the $(100-\alpha)\%$ CI for $(\mu_1 - \mu_2)$ is: MEAN W 2 IND. SAMPLES

KNOWN VARIANCES
· known population variances σ_1^2 & σ_2^2 $\quad (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

UNKNOWN VARIANCES
· unknown population variances $\quad (\bar{x}_1 - \bar{x}_2) \pm t_{\nu,\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
· Sample variances s_1^2, s_2^2

Have?	Sample 1	Sample 2
size	n_1	n_2
mean	$\bar{x}_1$	$\bar{x}_2$

Where t has $\nu = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1-1)} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2-1)}}$ degrees of freedom (round to nearest int)

Before & After — MEAN W 2 PAIRED SAMPLES

- Then the $(100-\alpha)\%$ CI for $\mu_d = (\mu_1 - \mu_2)$ is: $\quad \bar{d} \pm t_{\nu,\alpha/2}\left(\frac{s_d}{\sqrt{n}}\right)$

Have?
· 2 PAIRED samples of size n, leading to a "difference sample". mean $\bar{d}$ SD s_d

Where t has $\nu = (n-1)$ degrees of freedom.

*LARGE sample: $np \geq 5$ & $n(1-p) \geq 5$ — ONE-SAMPLE PROPORTION

Then the $(100-\alpha)\%$ CI for p is: $\quad \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Have?	Sample
size	n
success	$\hat{p}$
failure	$1-\hat{p}$

TWO-SAMPLE DIFFERENCE IN PROPORTIONS

Then the $(100-\alpha)\%$ CI for $(p_1 - p_2)$ is: $\quad (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Have?	Sample 1	Sample 2
size	n_1	n_2
success	$\hat{p}_1$	$\hat{p}_2$

*$n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$, $n_2(1-\hat{p}_2) \geq 5$

VARIANCE W 1 SAMPLE

- Then the $(100-\alpha)\%$ CI for σ^2 is (note the asymmetry!):

$$\frac{(n-1)s^2}{X^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{X^2_{1-\alpha/2}}$$

Have?	Sample
size	n
variance	s^2

Where both X^2 values have $\nu = n-1$ degrees of freedom.

*normally distributed populations — VAR W 2 SAMPLES

- Then the $(100-\alpha)\%$ CI for σ_1^2/σ_2^2 is:

$$\frac{s_1^2}{s_2^2 f_{\alpha/2}(\nu_1,\nu_2)} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2 f_{\alpha/2}(\nu_2,\nu_1)}{s_2^2}$$

Have?	Sample 1	Sample 2
size	n_1	n_2
variance	s_1^2	s_2^2

With degrees of freedom $\nu_1 = (n_1 - 1)$ and $\nu_2 = (n_2 - 1)$.

PREDICTION INTERVAL

*Normally distributed data

- Then the $(100-\alpha)\%$ PI for a future observation x_0 is:

KNOWN POP VAR
· Population variance σ^2 $\quad \bar{x} \pm z_{\alpha/2}\sigma\sqrt{1+1/n}$

UNKNOWN POP VAR
· Sample variance s^2 $\quad \bar{x} \pm t_{\alpha/2}s\sqrt{1+1/n}$

Have?	Sample
size	n
mean	$\bar{x}$

With t having $\nu = (n-1)$ degrees of freedom.

ESTIMATES

To estimate some population parameter θ, we want to find a **point estimate** $\hat{\theta}$, which is a single value of a statistic $\tilde{\Theta}$.

A statistic $\tilde{\Theta}$ is an **unbiased estimator** of the parameter θ if:

$$\mu_{\tilde{\Theta}} = E[\tilde{\Theta}] = \theta$$

Binomial Experiment

p - population success proportion

$\hat{p} = \frac{X}{n}$ - Sample success proportion

$\mathbb{P}$ - statistic = $\{\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_n\}$

- An **interval estimate** of parameter θ is an interval

$$\hat{\theta}_L \leq \theta \leq \hat{\theta}_U$$

where $\hat{\theta}_L, \hat{\theta}_U$ depend on the statistic $\tilde{\Theta}$ being sampled as well as its sampling distribution.

- The interval $[\hat{\theta}_L, \hat{\theta}_U]$ is referred to as the $\mathbf{100(1-\alpha)\%}$ **confidence interval** (CI) = centrality point estimate ± spread point estimate

ERROR TYPES

- Type I error: reject H_0 when it is true — to reduce α: ↑ sample size, ↓ critical region size
  - P(committing Type I error) = α
  - α = level of significance
- Type II error: fail to reject (FTR) H_0 when it is false — to reduce β: ↑ sample size, ↑ critical region size
  - P(committing Type II error) = β
  - We can only find β if we have a specific H_A
  - Statistical Power: the probability of correctly rejecting H_0 = $(1-\beta)$

In reality, Ho is:

Our Decision: H_0	True	False
Reject	Type I error (α)	Power ($1-\beta$)
Do Not Reject	Confidence level ($1-\alpha$)	Type II error (β)

a hypothesis is a conjecture of 1+ populations

TESTING PROCEDURE 1 V 2 SIDED

Hypothesis testing procedure

1. Come up with a **null hypothesis (H_0)**: one that captures the belief we are prepared to accept in absence of data
2. Come up with an **alternate hypothesis (H_A)**: one that we would accept if the null hypothesis is false
3. Based on test statistic, reject or FTR the null hypothesis

A hypothesis is **2-sided** if it specifies that the parameter is exactly equal to a given value
- The mean height of MIE237 students is 175 cm
- The variance of MIE237 student heights is 10 cm

A hypothesis is **1-sided** if it specifies that the parameter value is at least OR at most a given value
- The mean height of MIE237 students is at most 177 cm
- The variance of MIE237 student heights is at least 10 cm

if p-value is greater α - FTR H_0 (not in rejection region)

if p-value is less than α - Reject H_0 (in rejection region)

To Enter SD Mode: MODE MODE 1 (SD)
Clear input data: SHIFT MODE 1 = (clear)(scl)
Type num then M+ to input values

HYPOTHESIS TESTS

one-sample z test — One-Sample

- The test statistic z is given as:
  · Known Population variance σ^2 $\quad z = \frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}}$

Have?	Sample
size	n
mean	$\bar{x}$

We reject H_0 if:
- H_0 is $\mu = \mu_o$ AND either $z > z^*_{1-\alpha/2}$ or $z < z^*_{\alpha/2}$
- H_0 is $\mu \leq \mu_o$ AND $z > z^*_{1-\alpha}$
- H_0 is $\mu \geq \mu_o$ AND $z < z^*_{\alpha}$

one-sample t-test

- The test statistic t is given as:
  · Sample variance s^2 $\quad t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$

With $\nu = n-1$ degrees of freedom.
We reject H_0 if:
- H_0 is $\mu = \mu_o$ AND either $t > t^*_{\alpha/2}$ or $t < -t^*_{\alpha/2}$
- H_0 is $\mu \leq \mu_o$ AND $t > t^*_{\alpha}$
- H_0 is $\mu \geq \mu_o$ AND $t < -t^*_{\alpha}$

two-sample z-test — Two-Sample

- Then the z statistic is:
  · Known Population variances σ_1^2 & σ_2^2 $\quad z = \frac{(\bar{x}_1 - \bar{x}_2) - d_o}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$

Have?	Sample 1	Sample 2
size	n_1	n_2
mean	$\bar{x}_1$	$\bar{x}_2$

We reject H_0 if:
- H_0 is $(\mu_1 - \mu_2) = d_o$ AND either $z > z^*_{1-\alpha/2}$ or $z < z^*_{\alpha/2}$
- H_0 is $(\mu_1 - \mu_2) \leq d_o$ AND $z > z^*_{1-\alpha}$
- H_0 is $(\mu_1 - \mu_2) \geq d_o$ AND $z < z^*_{\alpha}$

two-sample t-test

- Then the t statistic is given as $t = \frac{(\bar{x}_1 - \bar{x}_2) - d_o}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$
  · Sample variances s_1^2 & s_2^2

Where t has $\nu = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1-1)} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2-1)}}$ degrees of freedom (round to nearest int)

We reject H_0 if one of these apply, and FTR otherwise
- Ho is $(\mu_1 - \mu_2) = d_o$ AND either $t > t^*_{\alpha/2}$ or $t < -t^*_{\alpha/2}$
- Ho is $(\mu_1 - \mu_2) \leq d_o$ AND $t > t^*_{\alpha}$
- Ho is $(\mu_1 - \mu_2) \geq d_o$ AND $t < -t^*_{\alpha}$

Paired t-test

- Then the t statistic with $\nu = (n-1)$ degrees of freedom is: $\quad t = \frac{\bar{d} - d_o}{s_d/\sqrt{n}}$

Have:
· Paired sample of size n
· mean difference $\bar{d}$
· SD s_d

We reject H_0 if one of these three apply; otherwise we FTR
- H_0 is $\bar{d} = d_o$ AND either $t > t^*_{\alpha/2}$ or $t < -t^*_{\alpha/2}$
- H_0 is $d \leq d_o$ AND $t > t^*_{\alpha}$
- H_0 is $d \geq d_o$ AND $t < -t^*_{\alpha}$

*all $e_i \geq 5$ — Chi-Squared goodness of fit test

- Then the X^2 statistic is: $X^2 = \sum_i^k \frac{(o_i - e_i)^2}{e_i}$ — Have: expected & observed, k bins in k bins, frequency per bin e_i & o_i
  - With $\nu = (k-1)$ degrees of freedom

We reject Ho if $X^2 > X^2_\alpha$ for a pre-selected α

Then the z statistic is given as: — One-Proportion z-test

* $n\hat{p}$ & $n(1-\hat{p})$ both ≥ 5

$$z = \frac{\hat{p} - p_o}{\sqrt{p_o(1-p_o)/n}}$$

Have?	Sample
size	n
success	$\hat{p}$
failure	$1-\hat{p}$

We reject H_0 if:
- H_0 is $p = p_o$ AND either $z > z^*_{1-\alpha/2}$ or $z < z^*_{\alpha/2}$
- H_0 is $p \leq p_o$ AND $z > z^*_{1-\alpha}$
- H_0 is $p \geq p_o$ AND $z < z^*_{\alpha}$

* $n\hat{p}$ & $n(1-\hat{p})$ both ≥ 5 for both samples — Two Proportion z-test

- Then the z statistic is given as:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{x_1 + x_2}{n_1 + n_2}\right)\left(1 - \frac{x_1 + x_2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Have?	Sample 1	Sample 2
size	n_1	n_2
success	$\hat{p}_1$	$\hat{p}_2$

$\hat{p}_1 = x_1/n_1$ $\quad \hat{p}_2 = x_2/n_2$

We reject H_0 if:
- H_0 is $p_1 = p_2$ AND either $z > z_{1-\alpha/2}$ or $z < z^*_{\alpha/2}$
- H_0 is $p_1 \leq p_2$ AND $z > z^*_{1-\alpha}$
- H_0 is $p_1 \geq p_2$ AND $z < z^*_{\alpha}$

Chi-Squared test for single variance

- Then the X^2 statistic with $\nu = (n-1)$ degrees of freedom is:

$$X^2 = \frac{(n-1)s^2}{\sigma_o^2}$$

Have?	Sample
size	n
variance	s^2
null hyp var	σ_o^2

We reject H_0 if (note the asymmetry!):
- H_0 is $\sigma^2 = \sigma_o^2$ AND either $X^2 < X^{2^*}_{1-\alpha/2}$ or $X^2 > X^{2^*}_{\alpha/2}$
- H_0 is $\sigma^2 \leq \sigma_o^2$ AND $X^2 > X^{2^*}_{\alpha}$
- H_0 is $\sigma^2 \geq \sigma_o^2$ AND $X^2 < X^{2^*}_{1-\alpha}$

F-test for ratio of variances

- Then the F statistic with degrees of freedom $\nu_i = (n_i - 1)$ is: $\quad f = \frac{s_1^2}{s_2^2}$

Have?	Sample 1	Sample 2
size	n_1	n_2
variance	s_1^2	s_2^2

We reject H_0 if:
- H_0 is $\sigma_1^2 = \sigma_2^2$ AND either $f < f^*_{1-\alpha/2}(\nu_1,\nu_2)$ or $f > f^*_{\alpha/2}(\nu_1,\nu_2)$
- H_0 is $\sigma_1^2 \leq \sigma_2^2$ AND $f > f^*_{\alpha}(\nu_1,\nu_2)$
- H_0 is $\sigma_1^2 \geq \sigma_2^2$ AND $f < f^*_{1-\alpha}(\nu_1,\nu_2)$

Chi-Squared of independance/homogeneity

The expected value of each cell is given as:

$$e_{ij} = \frac{(\text{row } i \text{ total})(\text{row } j \text{ total})}{\text{grand total}}$$

*all $e_{ij} \geq 5$ for all i, j

- The X^2 statistic for independence is given as:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Have: expected & observed frequency of cell e_{ij} & o_{ij}

With X^2 having $\nu = (|I|-1)(|J|-1)$ degrees of freedom.
We reject H_0 if $X^2 > X^{2^*}_{\alpha}$

- Special case: if $|I| = |J| = 2$ (i.e., a 2x2 table), then: — YATES CORRECTION

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(|o_{ij} - e_{ij}| - 0.5)^2}{e_{ij}}$$

To recall this type of value:	Perform this key operation:
Σx^2	SHIFT 1 (S-SUM) 1 (Σx^2)
Sum of values Σx	SHIFT 1 (S-SUM) 2 (Σx)
# of data n	SHIFT 1 (S-SUM) 3 (n)
mean $\bar{x}$	SHIFT 2 (S-VAR) 1 ($\bar{x}$)
Population SD σ_x	SHIFT 2 (S-VAR) 2 (σ_x)
Sample SD s_x	SHIFT 2 (S-VAR) 3 (s_x)

HYPOTHESIS TESTING

STATISTICS — LINEAR REGRESSION

ANOVA

Connection to experiment design!

SSR = sum of squares (regression) = $\sum(\hat{y}_i - \bar{y})^2 = b_1 S_{XY}$
- Amount of error incorporated by the linear regression model as oppose to simply fitting the mean value of y

SSE = sum of squares (error) = $\sum(y_i - \hat{y}_i)^2$
- Amount of error between the regression model and actual data

SST = sum of squares (total) = $\sum(y_i - \bar{y})^2 = SSR + SSE = S_{YY}$
- Amount of total error in the data if fitting to the mean value of y

Source of Variation	Sum of Squared Error	Degrees of Freedom	Mean Squared Error	F statistic
Regression	SSR	1	$\frac{SSR}{1} = SSR$	$\frac{SSR}{s^2} = f_{(v_1=1, v_2=n-2)}$
Error	SSE	$n-2$	$\frac{SSE}{n-2} = s^2$	
Total	SST	$n-1$		

- We reject H_o: $\beta_1 = 0$ if $f > f^*_{(v_1=1, v_2=n-2, \alpha)}$
  - Upper-tailed: the rejection region is only if $f > f^*$ (so α is not halved)

LR MODEL

The general form of a simple linear regression is:

Note: Y & ϵ are RV
*ϵ is normally distributed w mean 0, variance σ^2

$Y = \underbrace{\beta_0}_{\text{intercept}} + \underbrace{\beta_1}_{\text{slope}} \underbrace{X}_{\text{input}} + \underbrace{\epsilon}_{\text{residual - diff btwn observed \& predicted value}}$

ORDINARY LEAST SQUARES

unconstrained, non-linear program

$$b_1 = \frac{n\sum_i(x_i y_i) - \sum_i(x_i)\sum_i(y_i)}{n\sum_i(x_i^2) - \left(\sum_i(x_i)\right)^2} = \frac{\sum_i(x_i - \bar{x})(y_i - \bar{y})}{\sum_i(x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}}$$

$$b_0 = \frac{\sum_i y_i - b_1 \sum_i x_i}{n} = \bar{y} - b_1\bar{x}$$

$$S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = (S_X)^2(n-1)$$

$$S_{YY} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (S_Y)^2(n-1)$$

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

$$s^2 = \sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n-2} = \frac{S_{YY} - b_1 S_{XY}}{n-2}$$

mean sq error

CALCULATOR

MODE MODE 2 1	enter linear regression mode
SHIFT CLR 1 (Scl) =	clear statistical memory
<x-data> , <y-data> DT (M+)	enter statistical data

To recall this type of value:	Perform this key operation:
Σx^2	SHIFT S-SUM 1
Σx	SHIFT S-SUM 2
n	SHIFT S-SUM 3
Σy^2	SHIFT S-SUM ▶ 1
Σy	SHIFT S-SUM ▶ 2
Σxy	SHIFT S-SUM ▶ 3
$\bar{x}$	SHIFT S-VAR 1
$x\sigma_n$	SHIFT S-VAR 2
$x\sigma_{n-1}$	SHIFT S-VAR 3
$\bar{y}$	SHIFT S-VAR ▶ 1
$y\sigma_n$	SHIFT S-VAR ▶ 2
$y\sigma_{n-1}$	SHIFT S-VAR ▶ 3
intercept → Regression coefficient A	SHIFT S-VAR ▶ ▶ 1 (B_0)
slope → Regression coefficient B	SHIFT S-VAR ▶ ▶ 2 (B_1)

B_1

CI
- The $100(1-\alpha)\%$ CI for β_1 in the model $Y = \beta_0 + \beta_1 x$ is:

$$b_1 \pm t_{\alpha/2}\frac{s}{\sqrt{S_{XX}}}$$

where t has $(n-2)$ degrees of freedom.
The t statistic for testing β_1 is:

$$t = \frac{b_1 - \beta_{1_o}}{s/\sqrt{S_{XX}}}$$

HYPO TEST
Where t has $v = (n-2)$ degrees of freedom.
We reject H_o if:
- H_o: $\beta_1 = \beta_{1_o}$ AND $t > t^*_{v,\alpha/2}$ or $t < -t^*_{v,\alpha/2}$
- H_o: $\beta_1 \le \beta_{1_o}$ AND $t > t^*_{v,\alpha}$
- H_o: $\beta_1 > \beta_{1_o}$ AND $t < -t^*_{v,\alpha}$

FTR β_1 suggests there is NO LINEAR relationship btwn x & y

B_0

CI
- The $100(1-\alpha)\%$ CI for β_0 is:

$$b_0 \pm t_{\alpha/2}\frac{s}{\sqrt{nS_{XX}}}\sqrt{\sum_i^n x_i^2}$$

Where t has $v = (n-2)$ degrees of freedom.
- The t statistic for testing β_0 is:

$$t = \frac{b_0 - \beta_{0_o}}{s\sqrt{\sum_i x_i^2/(nS_{XX})}}$$

HYPO TEST
Where t has $v = (n-2)$ degrees of freedom.
We reject H_o if:
- H_o: $\beta_1 = \beta_{1_o}$ AND $t > t^*_{v,\alpha/2}$ or $t < -t^*_{v,\alpha/2}$
- H_o: $\beta_1 \le \beta_{1_o}$ AND $t > t^*_{v,\alpha}$
- H_o: $\beta_1 > \beta_{1_o}$ AND $t < -t^*_{v,\alpha}$

Rejecting H_o implies the linear model can explain a significant portion of variability in y

Pop. Correlation Coefficient

$\rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$

Bounded btwn [-1,+1]
Dimensionless

SAMPLE CORRELATION COEF

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

$|r| \to 0$, weak assoc. btwn X&Y
$|r| \to 1$, strong assoc. btwn X&Y
+r: increasing y_i as $x \to \infty$
-r: decreasing y_i as $x \to \infty$

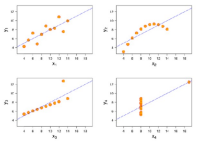
COEF. OF DETERMINATION

$$r^2 = \left(\frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}\right)^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Bounded by [0,1]
dimensionless
proportion of variability explainable by linear model

"correlation does not equal causation"
Just b/c the model fits, does NOT mean that ↑x results in definite change in Y.
LR can be a causal model but isn't always one

Anscombe's quartet is a good counterexample for why we should visualize & NOT rely soley on metrics
All 4 plots on the right have the same:
- Linear regression Line $\hat{y} = 3 + 0.5x$
- Sample variances S_{XX} & S_{YY}
- Correlation r & R^2



LOGISTIC REG MODEL

Logistic regression function:

$$Y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}$$

Our fitted model then becomes:

$$\hat{y} = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + \cdots + b_k x_k)}}$$

- Let's define the **odds of success** as $\frac{\hat{p}}{1-\hat{p}}$
- Then the **odds ratio** of a change from $\hat{p}_1$ to $\hat{p}_2$ is defined as

$$\frac{\hat{p}_2/(1-\hat{p}_2)}{\hat{p}_1/(1-\hat{p}_1)}$$

This means that as we change from $\hat{p}_1$ to $\hat{p}_2$, the *odds* of success changes by a factor of $\frac{\hat{p}_2/(1-\hat{p}_2)}{\hat{p}_1/(1-\hat{p}_1)}$

Example

Let's say we are trying to predict whether students pass a course based on how much they study.

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

Our model is:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1(\text{Hours}))}}$$

- Now let's run a logistic regression in R:

$H_o: \beta_1 = 0$
$H_a: \beta_1 \neq 0$

- This shows us that:
  - $b_0 = -4.0777$
  - $b_1 = 1.5046$
- Note that we don't have a specific F statistic for this output, but our individual coefficients are significant at the 0.05 level

```
> LogReg = glm(pass ~ hours, family=binomial)
> summary(LogReg)

Call:
glm(formula = pass ~ hours, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0777     1.7610  -2.316   0.0206 *
hours         1.5046     0.6287   2.393   0.0167 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.725  on 19  degrees of freedom
Residual deviance: 16.060  on 18  degrees of freedom
AIC: 20.06
```

as (#hours) ↑ 1, $\hat{p}$(pass) ↑ $\frac{1}{1+e^{-1.5}}$

OUTLIER ANAYLSIS

SD of residuals
$s_r = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$

Standardized res.
$SR_i = \frac{e_i}{s_r}$
if $|SR_i| > 3$, i could be an outlier

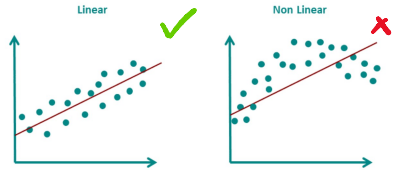
Studentized res.
$StR_i = \frac{e_i}{s_{r_i}^*}$ ← SD of residuals w/o point i
if $|StR_i| > 3$, i could be an outlier

Causes of outliers:

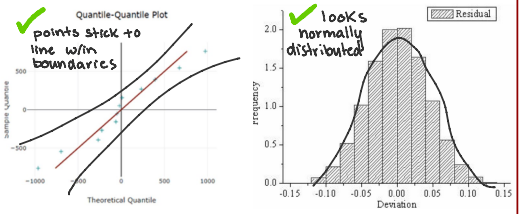
- Measurement error — REMOVE + RERUN LR
- Extreme but part of normal process — KEEP
- Extreme and *not* part of normal process — REMOVE + RERUN LR

Assumptions of LR

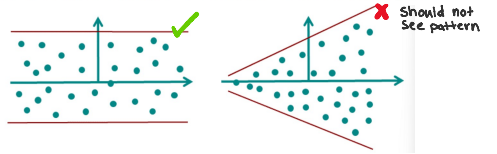
1. Linearity → scatterplot of x vs y



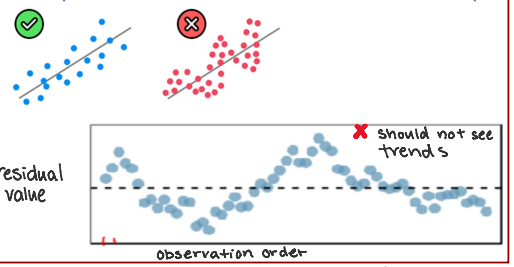
2. Normality of residuals → histogram of residuals, Normal QQ plot



3. Constant variance → residual vs. fitted value plot



4. Independence of errors → residual vs. observation order plot



MIE 247

MULTIPLE LINEAR REG

The multiple linear regression model is given as:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

Where we now have k many explanatory variables and therefore k-many β coefficients

*explanatory variables have power of 1 & no interaction w/ each other

ORDINARY LEAST SQUARES

- Let the general MLR equation be:

$$\mathbf{y} = \mathbf{b}^T\mathbf{X} + \mathbf{e}$$

Where $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{k,1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & \cdots & x_{k,n} \end{bmatrix}$, $\mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$

Then the solution of the point estimates vector **b** is:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$s^2 = \frac{SSE}{n-k-1} = \frac{\sum_i^n (y_i - \hat{y}_i)^2}{n-k-1}$$ variance of residuals

CI

- The $100(1-\alpha)\%$ CI for β_i is:

$$b_i \pm t_{v,\alpha/2}\sqrt{s^2 c_{ii}}$$

Where t has $v = n - k - 1$ degrees of freedom.

β_i

The t statistic with $v = n - k - 1$ degrees of freedom is:

$$t = \frac{b_i - \beta_{i_0}}{s\sqrt{c_{ii}}}$$

HYPO TEST

We reject H_o if:

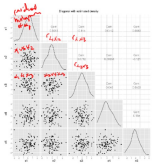
- H_o: $\beta_i = \beta_{i_0}$ AND $t > t^*_{v,\alpha/2}$ or $t < -t^*_{v,\alpha/2}$
- H_o: $\beta_i \leq \beta_{i_0}$ AND $t > t^*_{v,\alpha}$
- H_o: $\beta_i > \beta_{i_0}$ AND $t < -t^*_{v,\alpha}$

ASSUMPTIONS OF MLR

1. Linearity
2. Normality of residuals
3. Constant variance
4. Independence of errors

However, now we have an additional assumption of:

5. No multicollinearity → matrix plot of correlations (correlogram)



avoid correlations > |0.7|

Correlation between x_i and x_j:

$$r_{x_i x_j} = \frac{S_{x_i x_j}}{\sqrt{S_{x_i x_i} S_{x_j x_j}}}$$

ANALYSIS OF VARIANCE APPROACH

SSR = sum of squares (regression) = $\sum(\hat{y}_i - \bar{y})^2$

SSE = sum of squares (error) = $\sum(y_i - \hat{y}_i)^2$

SST = sum of squares (total) = $\sum(y_i - \bar{y})^2 = SSR + SSE$

Source of Variation	Sum of Squared Error	Degrees of Freedom	Mean Squared Error	F statistic
Regression	SSR	k	$\frac{SSR}{k} = MSR$	$\frac{MSR}{MSE} = f_{(v_1=k, v_2=n-k-1)}$
Error	SSE	$n-k-1$	$\frac{SSE}{n-k-1} = MSE$	
Total	SST	$n-1$		

H_o: $\beta_1 = \cdots = \beta_k = 0$

H_A: At least one $\beta_i \neq 0$

We reject H_o if $f > f^*_{v_1,v_2,\alpha}$

```
> summary(Full_Model)

Call:
lm(formula = y ~ x1 + x2 + x3, data = d)

Residuals:
Min      1Q      Median  3Q     Max
-1.8532 -1.4495 -0.3219 0.5919 3.2121

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  b0       5.8871     t0      9.36e-05 ***
x1           b1       0.1909     t1      0.000479 ***
x2           b2       0.2673     t2      6.58e-05 ***
x3           b3       0.6171     t3      0.591572
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error:  S  on 9 degrees of freedom
Multiple R-squared: 0.9117, Adjusted R-squared: 0.8823
F-statistic: f-value on K or v1 and V2 DF, p-value: 4.496e-05
```

Estimate | Standard Error | t-value

QUALITY OF FIT & MODEL SELECTION

As the number of fitted explanatory variables goes up, $R^2 \to 1$

- This is because the value of $SSE = \sum(y_i - \hat{y}_i)^2$ will decrease, while $SST = \sum(y_i - \bar{y})^2$ remains constant

The **Adjusted R^2** value divides the terms in R^2 (i.e., adjusts) by their degrees of freedom:

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

Factors to consider:

- Higher R^2_{adj} > lower R^2_{adj}
- Simpler models > more complicated models

if ↓ in SSE doesn't match w the ↓DOF, R^2_{adj} decreases to indicate poorer fit

INTERACTION TERMS

Models that contain different types of interaction terms

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3$$

If main effect terms are not significant, then unlikely interaction terms will be either

- Analyze partial models first and then add interaction terms

a) α = P(reject H_o | H_o is true) = P($\bar{x} < 14.9 \mid \mu \geq 15$)
= P($z < \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$) = P($z < \frac{14.9-15}{0.5/\sqrt{50}}$) = P($z < -1.41$)
From Normal table, $P(z < -1.41) = 0.0793 = \alpha$

b) β = P(fail to reject H_o | H_o is false) = P($\bar{x} > 14.9 \mid \mu_A = 14.8$)
= P($z > \frac{\bar{x}-\mu_A}{\sigma/\sqrt{n}}$) = P($z > \frac{14.9-14.8}{0.5/\sqrt{50}}$) = P($z > 1.41$)
From Normal table, $P(z > 1.41) = 1 - P(z < 1.41) = 1 - 0.9207$
$= 0.0793 = \beta$

CATAGORICAL VARIABLES

Add indicator or dummy variables

- Example:

$$z_A = \begin{cases} 1 \text{ if Yeast A} \\ 0 \text{ otherwise} \end{cases} \qquad z_B = \begin{cases} 1 \text{ if Yeast B} \\ 0 \text{ otherwise} \end{cases}$$

If adding indicator variables:

- Indicator main effect will affect the intercept values
- Indicator interaction effect will affect the slope values

A bakery is interested in modelling the amount of bread baked using the pH of the water and one of three types of yeast.

If we use Yeast A, our model is:

$\hat{y}_A = -161.897 + 54.294(\text{pH}) + 89.998 + 0$
$= -71.899 + 52.294(\text{pH})$

If we use Yeast B, our model is:

$\hat{y}_B = -161.897 + 54.294(\text{pH}) + 0 + 27.166$
$= -134.731 + 52.294(\text{pH})$

If we use Yeast C, our model is:

$\hat{y}_C = -161.897 + 54.294(\text{pH}) + 0 + 0$
$= -161.897 + 54.294(\text{pH})$

```
> Model2 = lm(y2 ~ ph + z1 + z2)
> summary(Model2)

Call:
lm(formula = y2 ~ ph + z1 + z2)

Residuals:
    Min      1Q  Median      3Q     Max
-31.038 -13.640   3.601  10.798  26.374

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -161.897     37.433  -4.325 0.000699 ***
ph            54.294      4.755  11.417 1.77e-08 ***
z1            89.998     11.052   8.143 1.11e-06 ***
z2            27.166     11.010   2.467 0.027127 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.05 on 14 degrees of freedom
Multiple R-squared:  0.9404,    Adjusted R-squared:  0.9277
F-statistic: 73.68 on 3 and 14 DF,  p-value: 8.14e-09
```

√n-k-1 | compare t to t* | reject Ho if α ≤ 0.001=α | f > f*(3, 14) | <0.05

ONE-FACTOR EXPERIMENT

We assume that each of the k populations are: ASSUMPTIONS
- Independent
- Have no outliers
- Normally distributed with individual mean μ_k and common variance σ^2

ANOVA statistical model: grand mean; deviation of level i observation j & mean of level i; if no stat effect (null) ↳ $\sum \alpha_i = 0$

value of level i observation j $\rightarrow y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ ← effect of level i

*ϵ_{ij} is normally distributed w/ mean 0 & variance σ^2

- In this model, we are then testing the hypothesis:
  - H_o: $\alpha_1 = \cdots = \alpha_k = 0$
    - In other words, none of the levels have an effect on the observed value
  - H_A: at least one $\alpha_i \neq 0$
    - In other words, at least one level has an effect on the observed value
- This is an upper-tailed F-test with significance level α
  - SSA = sum of squares (effects) = $\sum_i n_i(\bar{y}_{i.} - \bar{y}_{..})^2$
  - SSE = sum of squares (error) = $\sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$
  - SST = sum of squares (total) = $\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = SSA + SSE$

Source of Variation	Sum of Squared Error	Degrees of Freedom	Mean Squared Error	F statistic
Effects	SSA	$k-1$	$\frac{SSA}{k-1} = MSA$	$\frac{MSA}{MSE} = f_{(v_1=k-1, v_2=N-k)}$
Error	SSE	$N-k$	$\frac{SSE}{N-k} = MSE$	Reject H_o if $F > F^*$
Total	SST	$N-1$		

If reject H_o: find which means don't match 3 ways

1. Aggregate groups using linear constrasts — PLANNED COMPARISONS (contrasts)

$w = \sum_i c_i \mu_i$ where $\sum_i c_i = 0$

indicate the 'side' the mean is on & strength of composition of gamp

2. Conduct Hypothesis Test H_o: $\sum_i c_i \mu_i = 0$; H_A: $\sum_i c_i \mu_i \neq 0$

Two contrasts $w_a = \sum_i c_i \mu_i$ and $w_b = \sum d_i \mu_i$ are orthogonal if and only if $\sum \frac{c_i d_i}{n_i} = 0$

$$SSw = \frac{(\sum_i c_i \bar{y}_{i.})^2}{\sum_i \left(\frac{c_i^2}{n_i}\right)} = \frac{\left[\sum_i \left(\frac{c_i Y_{i.}}{n_i}\right)\right]^2}{\sum_i \left(\frac{c_i^2}{n_i}\right)}$$

Sum of sq's for contrast w

$\bar{y}_{i\cdot}$ mean; $Y_{i\cdot}$ total; n_i size of i; c_i contrast coef of i

$$MSE = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2}{N-k} = \frac{SSE}{N-k}$$

$f_w = \frac{SSw}{MSE}$ We reject H_o if $f > f^*_{(v_1=1, v_2=N-k)}$ F-test for contrast w

If can guess which means are going to differ in advance — a priori PAIRWISE COMPARISON (t-tests)

can use pairwise ind. 2-sample t-tests

$H_{o(i,j)}$: $\mu_i = \mu_j$ or alternatively $\mu_i - \mu_j = 0$

$H_{A(i,j)}$: $\mu_i \neq \mu_j$ or alternatively $\mu_i - \mu_j \neq 0$

* EXPERIMENT-WISE TYPE I ERROR RATE

$1 - (1-\alpha)^r$, for r-many pairwise comparisons

as # of pairwise comparisons ↑ prob of committing type I error ↑

If only have posterior knowledge of significant differences — TUKEY'S TEST (posterior pairwise comparisons)

H_o: $\mu_i = \mu_j$

H_A: $\mu_i \neq \mu_j$

$$SRSE_{i,j} = \sqrt{\frac{MSE}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

$$q_{ij} = \frac{|\bar{y}_{i.} - \bar{y}_{j.}|}{SRSE_{ij}}$$

- Recall that H_o: μ
- If $q > q^*$, then we reject H_o
- If $q \leq q^*$, then we fail to reject H_o

Compare q to q^* from Tukey's Table

ONE-WAY ANOVA W/ REPEATED MEASURES

- SSA: sum of squares (effect) = $b\sum_i(\bar{y}_{i.} - \bar{y}_{..})^2$
- SSB: sum of squares (between subjects) = $k\sum_j(\bar{y}_{.j} - \bar{y}_{..})^2$
- SSE: sum of squares (error) = $\sum_i \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$
- SST: sum of squares (total) = $\sum_i^k \sum_j^s (y_{ij} - \bar{y}_{..})^2 = SSA + SSB + SSE$

Source of Variation	Sum of Squared Error	Degrees of Freedom	Mean Squared Error	F statistic
Effects	SSA	$k-1$	$\frac{SSA}{k-1} = MSA$	$\frac{MSA}{MSE} = f_{(v_1=k-1, v_2=(k-1)(b-1))}$
Subjects	SSB	$b-1$	$\frac{SSB}{b-1} = MSB$	
Error	SSE	$(k-1)(b-1)$	$\frac{SSE}{(k-1)(b-1)} = MSE$	Reject H_o if $F > F^*$
Total	SST	$kb-1$		

- H_o: $\sigma_1^2 = \cdots = \sigma_k^2$
- H_A: at least one variance is not equal

BARTLETT'S TESTS ↳ test for homogeneity of variance

$$s^2_{pooled} = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)s_i^2$$

$$b = \frac{1}{s^2_{pooled}}\left[(s_1^2)^{n_1-1} \times \cdots \times (s_k^2)^{n_k-1}\right]^{\frac{1}{N-k}}$$

Compare b to $b_k(\alpha, n_1, \ldots, n_k)$, where $b_k(\alpha, n_1, \ldots, n_k) \approx \frac{\sum_{i=1}^k n_i b_k(\alpha, n_i)}{N}$ and $b_k(\alpha, n_i)$ is from the Bartlett table

RANDOMIZED COMPLETE BLOCK DESIGN

- SSA: sum of squares (effect) = $b\sum_i(\bar{y}_{i.} - \bar{y}_{..})^2$
- SSB: sum of squares (blocks) = $k\sum_j(\bar{y}_{.j} - \bar{y}_{..})^2$
- SSE: sum of squares (error) = $\sum_i \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$
- SST: sum of squares (total) = $\sum_i^k \sum_j^s (y_{ij} - \bar{y}_{..})^2 = SSA + SSB + SSE$

Source of Variation	Sum of Squared Error	Degrees of Freedom	Mean Squared Error	F statistic
Treatment	SSA	$k-1$	$\frac{SSA}{k-1} = MSA$	$\frac{MSA}{MSE} = f_{(v_1=k-1, v_2=(k-1)(b-1))}$
Blocks	SSB	$b-1$	$\frac{SSB}{b-1} = MSB$	$\frac{MSB}{MSE} = f_{(v_1=b-1, v_2=(k-1)(b-1))}$
Error	SSE	$(k-1)(b-1)$	$\frac{SSE}{(k-1)(b-1)} = MSE$	Reject H_o if $F > F^*$
Total	SST	$kb-1$		

Completely Randomized design
Randomly assign subjects into any of the K treatments being analyzed

Randomized Block Design
used when there is a second factor to stratify by
1. stratify by factor B into blocks
2. randomize subjects into treatments along factor A

Randomized complete block design
assign each treatment once to each block, same as two factor experiment w/o interaction

Interpolation v. Extrapolation

Interpolation: estimating a value w/in the domain of data points
ex: predicting an intermediate value in LR

Extrapolation: estimating a value beyond the data's domain
ex: LR beyond the data

Two contrasts $w_a = \sum_i c_i \mu_i$ and $w_b = \sum d_i \mu_i$ are **orthogonal** if and only if $\sum_i \frac{c_i d_i}{n_i} = 0$

Orthogonality means that both SSw_a and SSw_b form independent components of SSA
- This means that we are separating SSA into their w_a and w_b "components"
- If w_a and w_b are not orthogonal, then their SSW_i will overlap

If there are k treatments, then there are at most $k-1$ orthogonal linear contrasts, where:

$$SSA = SSw_1 + SSw_2 + \cdots + SSw_{k-1}$$

PLANNED COMPARISONS

TWO-FACTOR EXPERIMENT W/ INTERACTION

*assume each of the a,b level combinations has exactly n observations

Same assumptions with a two-way ANOVA, but with ab groups:
- Independence of observations
- Equality of variances
- No outliers
- Normality

Hypotheses

Main effect of factor A:
- $H_{o(A)}$: $\alpha_1 = \cdots = \alpha_a = 0$
- $H_{A(A)}$: at least one α_i is not equal

Main effect of factor B:
- $H_{o(B)}$: $\beta_1 = \cdots = \beta_b = 0$
- $H_{A(B)}$: at least one β_j is not equal

Interaction effect of AB:
- $H_{o(AB)}$: $(\alpha\beta)_{1,1} = \cdots = (\alpha\beta)_{a,b} = 0$
- $H_{A(AB)}$: at least one $(\alpha\beta)_{i,j}$ is not equal

ANOVA MODEL

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

let α_i represent effects of A
let β_j represent effect of B

Constraints

$\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, and $\sum_i \sum_j (\alpha\beta)_{ij} = 0$

$$SSA = bn\sum_i(\bar{y}_{i..} - \bar{y}_{...})^2$$

$$SSB = an\sum_j(\bar{y}_{.j.} - \bar{y}_{...})^2$$

$$SS(AB) = n\sum_i\sum_j(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$$

$$SSE = \sum_i\sum_j\sum_k(y_{ijk} - \bar{y}_{ij.})^2$$

$$SST = \sum_i\sum_j\sum_k(y_{ijk} - \bar{y}_{...})^2 = SSA + SSB + SS(AB) + SSE$$

Source of Variation	Sum of Squared Error	Degrees of Freedom	Mean Squared Error	F statistic
Main Effect A	SSA	$a-1$	$\frac{SSA}{a-1} = MSA$	$\frac{MSA}{MSE} = f_{A(v_1=(a-1), v_2=ab(n-1))}$
Main Effect B	SSB	$b-1$	$\frac{SSB}{b-1} = MSB$	$\frac{MSB}{MSE} = f_{B(v_1=(b-1), v_2=ab(n-1))}$
Interaction	$SS(AB)$	$(a-1)(b-1)$	$\frac{SS(AB)}{(a-1)(b-1)} = MS(AB)$	$\frac{MS(AB)}{MSE} = f_{AB(v_1=(a-1)(b-1), v_2=ab(n-1))}$
Error	SSE	$ab(n-1)$	$\frac{SSE}{ab(n-1)} = MSE$	
Total	SST	$abn-1$		

Assess each H_o independently based on its f-statistic REJECT H_o if $f > f^*_{(v_1, v_2)}$

INTERACTION PLOTS

- Main effects:
  - Horizontal lines: evidence of no effect
  - Lines with a trend: evidence of effect
- Interactions:
  - Parallel lines: evidence of no effect
  - Intersecting lines: evidence of effect