tags: Methods1

NOTE:

field: 100000

field: Epidemiology Definition of Causation

field: Factor/variable X **causes** result Y if some cases of Y would not have occurred if X had been absent.

NOTE:

field: 100001

field: Sample variance

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

NOTE:

field: 100002

field: Population(s) of interest

field: The group to which you would like your answer to apply

NOTE:

field: 100003

field: Variable of Interest

field: A measurement that can be made on each individual/member of the population

NOTE:

field: 100004

field: Facts about Normal Distributions

field:

- If Z has a Normal(0,1) distribution then $X = \sigma Z + \mu$ has a Normal(μ, σ^2) distribution
- If X has a Normal(μ, σ^2) distribution, then $Z = \frac{X-\mu}{\sigma}$ has a Normal(0,1) distribution.
- If X has a Normal(μ_x, σ_x^2) distribution, and Y has a Normal(μ_y, σ_y^2) distribution, and X and Y are independent of each other, then $X + Y \sim$ Normal$(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$

NOTE:

field: 100005

field: Sample mean

field: $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$

NOTE:

field: 100006

field: Sampling distribution for population $Y \sim$ Normal(μ, σ^2)

field: $N(\mu, \sigma^2/n)$

NOTE:

field: 100007

field: Variance (Expected value)

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

NOTE:

field: 100008

field: Covariance

field: $Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$

NOTE:

field: 100009

field: If X and Y are independent (covariance)

field: The covariance is 0

NOTE:

field: 100010

field: If $Cov(X, Y) = 0$, (independence)

field: Cannot say that X and Y are independent

NOTE:

field: 100011

field: $Cov(X, X) =$

field: $Var(X)$

NOTE:

field: 100012

field: $X \sim N(\mu, \sigma^2)$

- $E(\bar{X}) =$
- $V(\bar{X}) =$

field:

- $E(\bar{X}) = \mu$
- $V(\bar{X}) = \sigma^2/n$

NOTE:

field: 100013

field: Central Limit Theorem (in words)

field: If the population distribution of a variable X has population mean μ and finite population variance σ^2, then the sampling distribution of the sample mean becomes closer and closer to a Normal distribution as the sample size n increases: $\bar{X} \sim N(\mu, \sigma^2/n)$

NOTE:

field: 100014

field: Central Limit Theorem (theoretical)

field: Let $X_1, X_2, \ldots X_n$ be an iid sample from some poupation distribution F with mean μ and variance $\sigma^2 < \infty$. Then as the sample size $n \to \infty$, we have

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to N(0, 1)$$

NOTE:

field: 100015

field: $X \sim (\mu, \sigma^2)$

- $E(\bar{X}) =$
- $V(\bar{X}) =$

field:

- $E(\bar{X}) = \mu$
- $V(\bar{X}) = \sigma^2/n$

NOTE:

field: 100016

field: Reject H_0 when H_0 True

field: Type I error (false positive)

NOTE:

field: 100017

field: Type I error (false positive)

field: Reject H_0 when H_0 True

NOTE:

field: 100018

field: Fail to Reject H_0 when H_0 false

field: Type II error

NOTE:

field: 100019

field: Type II error

field: Fail to Reject H_0 when H_0 false

NOTE:

field: 100020

field: Significance level

field: α the probability of a Type I error

NOTE:

field: 100021

field: Power (at θ_1)

field: Probability of rejecting the null hypothesis when θ_1 is the truth

NOTE:

field: 100022

field: Test for data setting: $X_1, X_2, \dots X_n$ iid with sample mean $\bar{X}$, and known population variance σ^2, Null hypothesis $\mu = \mu_0$

- Test name
- Test Statistic
- Test Reference Distribution
- Critical Value
  - Lower
  - Upper
  - Two sided
- Confidence interval
- pvalue
  - upper:
  - lower:
  - two-sided
- Consistent/Finite Sample Exact/ Asymptotically Exact

field: z-test

- Test statistic: $Z(\mu_0) = \frac{\bar{X}-\mu_0}{\sqrt{\sigma^2/n}}$
- Reference Distribution: Under $H_0, Z(\mu_0) \sim N(0,1)$
  - Lower: Reject when $Z(\mu_0) < z_\alpha$ = qnorm(α)
  - Upper: Reject when $Z(\mu_0) > z_{1-\alpha}$ =qnorm(1-α)
  - Two sided: Reject when $|Z(\mu_0)| > z_{1-\alpha/2}$ = qnorm(1 - $\alpha/2$)
- Confidence interval: $\bar{X} \pm z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}$
- pvalue:

- upper: 1 - $\Phi(z)$ = 1 - pnorm(z)
- lower: $\Phi(z)$ = pnorm(z)
- two-sided: $2(1 - \Phi(|z|))$ = 2*(1 - pnorm(abs(z)))

- Consistent: Yes /Finite Sample Exact: Yes if $X_i \sim N$/ Asymptotically Exact: Yes

NOTE:

field: 100023

field: Exactness (finite/asymptotic)

field: Under any setting for which the null hypothesis is true, is the actual rejection probability equal to the desired level α?

- Finite Sample Exact: for sample size n is $P(RejectH_0) = \alpha$ when H_0 is true?
- Asymptotic Exactness: As $n \to \infty$ does $P(RejectH_0) \to \alpha$ when H_0 is true?

NOTE:

field: 100024

field: When is a test exact?

field:

- A test is FSE if the reference distribution is the true distribution of the test statistic T when H_0 is true
- A test is AE if the reference distribution is the asymptotic distribution of the test statistic when H_0 is true.
- (Distribution of p-values should be Unif(0.1))

NOTE:

field: 100025

field: Consistency

field: When H_0 is false (the alternative hypothesis is true), does the rejection probability (probability reject the null) tend to 1 as $n \to \infty$?

NOTE:

field: 100026

field: Interpretation of Confidence intervals

field: $(1-\alpha)100\%$ of the time, intervals constructed in this manner will include μ

NOTE:

field: 100027

field: Test for data setting: $X_1, X_2, \ldots X_n$ iid with sample mean $\bar{X}$, and unknown population variance , Null hypothesis $\mu = \mu_0$

- Test name
- Test Statistic
- Test Reference Distribution
- Critical Value/ Rejection region
  - upper:
  - lower:
  - two-sided
- Confidence interval

- pvalue
  - upper:
  - lower:
  - two-sided
- Consistent/Finite Sample Exact/ Asymptotically Exact

field:

- Test name: t-test
- Test Statistic: $t(\mu_0) = \frac{\bar{X}-\mu_0}{\sqrt{s^2/n}}$
- Test Reference Distribution: t_{n-1}
- Critical Value/ Rejection region
  - upper: Reject if $t(\mu_0) > t_{(n-1),1-\alpha}$ = qt(1 - α,n-1)
  - lower: Reject if $t(\mu_0) < t_{n-1,\alpha}$
  - two sided: Reject if $|t(\mu_0)| > t_{n-1,1-\alpha/2}$
- Confidence interval: $\bar{X} \pm t_{n-1,1-\alpha/2}\sqrt{\frac{s^2}{n}}$
- pvalue, with $t(\mu_0) = t$, and pt representing the cdf of a t distribution
  - upper: 1 - pt($t, n-1$)
  - lower: pt(t,n-1)
  - two-sided: 2*(1 - pt(abs(t)),n-1)
- Consistent Yes/Finite Sample Exact Yes if normal/ Asymptotically Exact Yes

NOTE:

field: 100028

field: Test for data setting $Y_1, \ldots, Y_n$ iid Bernoulli(p) (Exact test option), parameter of interest p

field: Test for data setting $Y_1, \ldots, Y_n$ iid Bernoulli(p), parameter of interest p

- Test name: Exact Binomial Test (uses the distribution of the sum of Bern(p) RVs)
- Test Statistic: $X = \sum_{i=1}^{n} Y_i = n\bar{Y}$
- Test Reference Distribution: Under H_0 Binomial(n, p_0)
- Critical Value/ Rejection region: Sometimes use randomized test
  - upper: Reject H_0 for $X \geq c$ for c such that $P(X \geq c) \leq \alpha$
  - lower: Reject H_0 for $X \leq c$ for c such that $P(X \leq c) \leq \alpha$
  - two-sided: Reject H_0 for $p_0(X) \leq c$ for c such that$P_{H_0}(p_0(X) \leq c) \leq \alpha$, where $p_0(X)$ is $P(X = x)$ under H_0
- Confidence interval: Values that are not rejected
- pvalue: Sum of the probabilities that are less than or equal to the observed value (under the null hypothesis)
- Consistent/Finite Sample Exact/ Asymptotically Exact

NOTE:

field: 100029

field: Test for data setting $Y_1, \ldots, Y_n$ iid Bernoulli(p)
Parameter of interest: p (Not FSE method)

field: Test for data setting $Y_1, \ldots, Y_n$, parameter of interest: p iid Bernoulli(p) (option 2)

- Test name: Binomial z-test (Use when $np_0 > 5$ and $n(1-p_0) > 5$)
- Test Statistic: $X = \sum_{i=1}^{n} = n\bar{Y}$, $\hat{p} = X/n$,

$$z(p_0) = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

  (score)

- Test Reference Distribution: Under H_0, Approximately $X \sim N(np_0, np_0(1-p_0))$ and $z(p_0) \sim N(0,1)$
- Critical Value/ Rejection region
  - upper: $z(p_0) > z_{1-\alpha}$
  - lower: $z(p_0) < z_{\alpha}$
  - two-sided: $|z(p_0)| > z_{1-\alpha/2}$
- Confidence interval: Uses wald interval (derived from t-test) (with $z_w(p_0) = \frac{\hat{p}-p_0}{\sqrt{\hat{p}(1-\hat{p})/n}}$) $\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- pvalue
  - upper: 1 - $\Phi(z(p_0))$ = 1 - pnorm($z(p_0)$)
  - lower: $\Phi(z(p_0))$ = pnorm(z)
  - two-sided: $2(1 - \Phi(|z(p_0)|))$ = 2*(1 - pnorm(abs(z)))
- Consistent: Yes/Finite Sample Exact: No/ Asymptotically Exact: Yes

NOTE:

field: 100030

field: Continuity correction for Binomial z-test

field: With $X \sim Binom(n, p)$, instead of $P(X \leq x)$, use $P(W \leq x + 1/2)$ where $W \sim N(np, np(1-p))$

NOTE:

field: 100031

field: Data Setting: $X_1, \ldots, X_n$, iid parameter of interest: M - median $H_0 : M = M_0$

- Test name:
- Test Statistic
- Test Reference Distribution
- Critical Value/ Rejection region
  - upper:
  - lower:
  - two-sided
- Confidence interval
- pvalue
  - upper:
  - lower:
  - two-sided
- If ties?
- Consistent/Finite Sample Exact/ Asymptotically Exact

field:

- Test name: Sign Test
- Test Statistic: $Y_i = I(X_i < M_0)$, $\hat{p}_{M_0} = \frac{\sum Y_i}{n}$ (proportion of observations less than or equal to hypothesized median)
- Test Reference Distribution: Normal distribution: with $p_0 = .5$
- Critical Value/ Rejection region: $z = \frac{\hat{p}_{M_0} - p_0}{\sqrt{p_0(1-p_0)/n}}$
  - upper: $z > z_{1-\alpha}$
  - lower: $z < z_\alpha$
  - two-sided: $|z| > z_{1-\alpha/2}$
- Confidence interval: cant use the binomial proportion CI Set of values of M_0 that wouldn't be rejected at level α

$$\left(\frac{n - z_{1-\alpha/2}\sqrt{n}}{2}\right)^{th} \text{ Smallest Observation, } \left(\frac{n - z_{1-\alpha/2}\sqrt{n}}{2}\right)^{th} \text{ Smallest Observation}$$

- pvalue (binomial test on proportion)
  - upper: 1 - $\Phi(z(p_0))$ = 1 - pnorm($z(p_0)$)
  - lower: $\Phi(z(p_0))$ = pnorm(z)
  - two-sided: $2(1 - \Phi(|z(p_0)|))$ = 2*(1 - pnorm(abs(z)))
- If there are ties: remove all observations equal to M_0, then test prop of observations $< M_0$ given not equal to M_0 is .5
- Consistent: yes/Finite Sample Exact: No / Asymptotically Exact: yes

NOTE:

field: 100032

field: Data Setting: $X_1, \ldots, X_n$, iid
parameter of interest: M - median $H_0 : M = M_0$ (under symmetry assumption)

- Test name:
- Assumptions
- Procedure:
- Test Statistic, Test Reference Distribution

field: Data Setting: $X_1, \ldots, X_n$, iid parameter of interest: M - median $H_0 : M = M_0$ (option 1)

- Test name: Wilcoxon signed-rank test
- Assumptions: symmetry - equivalently a test of the mean. otherwise tests the pseudo-median
- Procedure: testing c_0 is the center (median)
  - Calculate distance of each observation from c_0
  - Rank observations by the distance (abs value) from c_0
- Test Statistic: S sum of the ranks that correspond to observations larger than c_0, $Z = \frac{S - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N(0,1)$
- Test Reference Distribution:
  - Exact p-value - assume each rank has the same chance of being assigned to observations above or below c_0 - all possible ways to assign the ranks
  - Normal approximation to the null distribution $S \sim N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$
- Consistent: Yes under symmetry assumption
  Finite Sample Exact: No
  Asymptotically Exact: Yes (under symmetry assumption)

NOTE:

field: 100033

field: Pseudomedian

field: Median of the distribution of sample means from samples of size 2

NOTE:

field: 100034

field: Data Setting: $X_1, \ldots, X_n$, iid $N(\mu, \sigma^2)$
Parameter of interest: $\sigma^2 = Var(X)$:
$H_0 : \sigma^2 = \sigma_0^2$

field:

- Test name: χ^2 for Population Variance
- Test Statistic $X(\sigma_0) = \frac{(n-1)s^2}{\sigma_0^2}$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- Test Reference Distribution: Under $H_0 : X(\sigma_0) = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}$
- Critical Value/ Rejection region
  - $\sigma^2 > \sigma_0^2$ Reject H_0 for $X(\sigma_0^2) > \chi^2_{n-1,(1-\alpha)}$
  - $\sigma^2 < \sigma_0^2$ Reject H_0 for $X(\sigma_0^2) < \chi^2_{n-1,(\alpha)}$
  - $\sigma^2 \neq \sigma_0^2$ Reject H_0 for $X(\sigma_0^2) > \chi^2_{n-1,(1-\alpha/2)}$ or $X(\sigma_0) < \chi^2_{n-1,(\alpha/2)}$
- Confidence interval
$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,(\alpha/2)}}\right)$$
- pvalue
  - $\sigma^2 > \sigma_0^2$: $p = 1 - pchisq(X(\sigma_0)^2, n-1)$
  - $\sigma^2 < \sigma_0^2 : p = pchisq(X(\sigma_0^2), n-1)$
  - $\sigma^2 \neq \sigma_0^2 : p = 2\min(1-pchisq(X(\sigma_0^2), n-1), pchisq(X(\sigma_0^2)), n-1)$
- (fill in later)Consistent?/Finite Sample Exact?/ Asymptotically Exact?

NOTE:

field: 100035

field: Data Setting: $X_1, \ldots, X_n$, iid
Parameter of interest: $\sigma^2 = Var(X)$,
Sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$,
$H_0 : \sigma^2 = \sigma_0^2$ (asymptotic)

- Test name:
- Test Statistic
- Test Reference Distribution
- Critical Value/ Rejection region
- Confidence interval
- pvalue

field: Data Setting: $X_1, \ldots, X_n$, iid parameter of interest: $\sigma^2 = Var(X)$, sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, $H_0 : \sigma^2 = \sigma_0^2$

- Test name: Asymptotic t-test for population variance
- Test Statistic: $Y = (X_i - \bar{X})^2$,

$$t(\sigma_0^2) = \frac{Y - \frac{\bar{n}-1}{n}\sigma_0^2}{\sqrt{s_y^2/n}} \to N(0,1)$$

  Note $\bar{Y} = \frac{n-1}{n}s^2$

- Tests that the population mean of the Y_i is $\frac{n-1}{n}\sigma_0^2$
- Test Reference Distribution $\frac{\frac{n-1}{n}s^2 - \frac{n-1}{n}\sigma^2}{\sqrt{Var(\frac{n-1}{n}s^2)}} = \frac{\bar{Y} - \frac{n-1}{n}\sigma^2}{\sqrt{Var(\bar{Y})}} \to N(0,1)$, so we can use t-test
- Critical Value/ Rejection region

- upper: Reject if $t(\sigma_0^2) > t_{(n-1),1-\alpha}$ = qt(1 - α,n-1)
- lower: Reject if $t(\sigma_0^2) < t_{n-1,\alpha}$
- two sided: Reject if $|t(\sigma_0^2)| > t_{n-1,1-\alpha/2}$

- Confidence interval: $\bar{X} \pm t_{n-1,1-\alpha/2}\sqrt{\frac{s^2}{n}}$
- pvalue, with $t(\mu_0) = t$, and pt representing the cdf of a t distribution
  - upper: 1 - pt($t, n-1$)
  - lower: pt(t,n-1)
  - two-sided: 2*(1 - pt(abs(t)),n-1)

NOTE:

field: 100036

field: Test for data setting $X_1, \ldots X_n$ iid from population distribution F. Test $H_0 : F = F_0$

- Test name:
- Process
- Test Statistic
- Test Reference Distribution
- Critical Value/ Rejection region

field: Test for data setting $X_1, \ldots X_n$ iid from population distribution F. Test $H_0 : F = F_0$

- Test name: Kolmogorov-Smirnov Test
- Process
- Test Statistic: $D(F_0) = sup_x|\hat{F}(x) - F_0(x)$, where $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x)$ is the empirical cdf and $F_0(x)$ is the null hypothesis cdf (maximum values of difference between emperical and null)

- Test Reference Distribution: Kolmogorov distribution
- Critical Value/ Rejection region: Reject for large values of $\sqrt{n}D(F_0)$
- Note the one sided version does not have an easy interpretation

NOTE:

field: 100037

field: Data setting: $X_1, \ldots, X_n$ iid from discrete distribution. Test fit of distribution

- Test name:
- Process
- Test Statistic
- Test Reference Distribution
- Critical Value/ Rejection region
- If parameter values of discrete distribution are not known

field: Data setting: $X_1, \ldots, X_n$ iid from discrete distribution. Test fit of distribution

- Test name: χ^2 goodness of fit test, test for discrete distributions
- Process: Test the underlying population distribution is $P(X = x) = p_0(x)$, where $\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i = x)$
  - Let $j = 1, \ldots, k$ the different categories that X_i can take
  - Let O_j be the observed number of observations that belong to category j
  - Let $E_j = np_0(j)$ be the expected number of observations that would belong to category j if the null hypothesis were true
- Test Statistic: $X(p_0) = \sum_x \frac{n(\hat{p}(x)-p_0(x))^2}{p_0(x)} = \sum_{j=1}^{k} \frac{(O_j-E_j)^2}{E_j}$

- Test Reference Distribution: Under H_0, $X(p_0) \to \chi^2_{k-1}$
- Critical Value/ Rejection region: Reject for large values of $X(p_0)$ - Reject H_0 for $X(p_0) > \chi^2_{k-1}(1-\alpha)$
- Note: Null hypothesis doesnt completely specify the distribution, just the family of distributions with perhaps unknown parameters
  - Estimate the parameters
  - Use null distribution with estimated parameter values for E_j
  - Compute χ^2 test statistic
  - Compare to χ^2_{k-d-1} distribution where k = number of categories, d = number of estimated parameters

NOTE:

field: 100038

field: Data setting $X_1, \ldots, X_n$, $Y_1, \ldots, Y_m$ iid with known σ_x, σ_y. Estimate $d = \mu_x - \mu_y$

field:

- Test name: 2 sample z test
- Test Statistic: $z(d_0) = \frac{(\bar{X}-\bar{Y})-d_0}{\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}}}$
- Test Reference Distribution: Under H_0, $z(d_0) \sim N(0,1)$
- Critical Value/ Rejection region
  - Lower: $d \leq d_0$ Reject when $z(d_0) < z_\alpha$ = qnorm(α)
  - Upper: $d \geq d_0$ Reject when $z(d_0) > z_{1-\alpha}$ =qnorm(1-α)
  - Two sided: $d \neq d_0$ Reject when $|z(d_0)| > z_{1-\alpha/2}$ = qnorm(1 - $\alpha/2$)
- Confidence interval:

$$(\bar{X} - \bar{Y}) \pm z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}$$

NOTE:

field: 100039

field: Data setting $X_1, \ldots, X_n$, $Y_1, \ldots, Y_m$ iid with unknown but equal σ_x, σ_y
Estimate $d = \mu_x - \mu_y$

- Test name:
- Estimate of $\sigma_x^2 = \sigma_y^2$
- When σ_x, σ_y not actually equal

field: Data setting $X_1, \ldots, X_n$, $Y_1, \ldots, Y_m$ iid with unknown σ_x, σ_y. Estimate d

- Test name: Equal variance 2-sample t-test
- Note: Estimate of $\sigma_x^2 = \sigma_y^2 = s_p^2 = \frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})}{(m-1)+(n-1)} = \frac{(m-1)s_x^2 + (n-1)s_y^2}{(m+n-2)}$ (weighted average of the two sample variances)
- Test Statistic: $t(d_0) = \frac{(\bar{X} - \bar{Y}) - d_0}{\sqrt{s_p^2(\frac{1}{m} + \frac{1}{n})}}$
- Test Reference Distribution: For Normal populations, under H_0: $t(d_0) \sim t_{m+n-2}$
- Critical Value/ Rejection region
  - $d > d_0$ Reject H_0 for $t_e(d_0) > t_{m+n-2,(1-\alpha)}$
  - $d < d_0$ Reject H_0 for $t_e(d_0) < t_{m+n-2,(\alpha)}$
  - $d \neq d_0$ Reject H_0 for $|t_e(d_0)| > t_{m+n-2,(1-\alpha/2)}$
- Confidence interval $(\bar{X} - \bar{Y}) \pm t_{m+n-2,(1-\frac{\alpha}{2})} \sqrt{s_p^2(\frac{1}{m} + \frac{1}{n})}$
- When not equal:
  - Expected value of Estimated variance is

    * larger than it should be when the smaller sample comes from the population with smaller variance
    * The test statistic will be closer to zero than it should be, and rejection rates will be smaller
    * Less power - more conservative
  - Expected value of Estimated variance is smaller than it should be
    * when smaller sample comes from the population with the larger variance
    * test statistic will have a larger absolute value than it should an rejection rates will be larger
    * more power - anti conservative

NOTE:

field: 100040

field: Data setting $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$ iid with unknown not equal σ_x, σ_y Estimate $d = \mu_x - \mu_y$

- Test name:
- Estimate of $Var(\bar{X} - \bar{Y})$
- Test Statistic
- Test Reference Distribution
- Degrees of freedom
- Critical Value/ Rejection region
- Confidence interval
- Compare to equal variance

field: Data setting $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ iid with unknown not equal equal σ_x, σ_y Estimate d

- Test name: Unequal variance 2 sample t-test
- Estimate of $Var(\bar{X} - \bar{Y}) = \frac{s_x^2}{m} + \frac{s_y^2}{n}$
- Test Statistic: $t_U(d_0) = \frac{(\bar{X}-\bar{Y})-d_0}{\sqrt{\frac{s_x^2}{m}+\frac{s_y^2}{n}}}$
- Test Reference Distribution: If the two distributions are Normal, there is not an exact distribution for the test statistic - Use Welch-Satterthwaite approximation: Estimate degrees of freedom

  $$v = \frac{\left(\frac{s_x^2}{m} + \frac{s_y^2}{n}\right)^2}{\frac{s_x^4}{m^2(m-1)} + \frac{s_Y^4}{n^2(n-1)}}$$

  $\min(m-1, n-1) \leq v \leq m+n-2$ Under H_0 $t_u(d_0)$ approx $\sim t_v$
- Critical Value/ Rejection region: same as t-test
- Confidence interval: $(\bar{X} - \bar{Y}) \pm t_v(1 - \frac{\alpha}{2})\sqrt{\frac{s_x^2}{m} + \frac{s_Y^2}{n}}$
- Compare to equal variance:
  - For unequal sample sizes with unequal population variances, equal variance t-test does not have correct calibration
  - When samples sizes are equal both test statistics are the same, but degrees of freedom differ
  - When equal variance assumption is true, equal variance has slightly better power, and very slightly better calibration (more exact)

NOTE:

field: 100041

field: Data setting $X_1, \ldots, X_n$ iid F_x,
$Y_1, \ldots, Y_n$ iid F_y,
X_i not independent Y_i,
$(X_1, Y_1), \ldots, (X_n, Y_n)$ iid F_{XY} $Cov(X_i, Y_i) = \sigma_{XY}$, $Cov(X_i, Y_j) = 0$.
Estimate $d = \mu_x - \mu_y$, when $\sigma_x^2, \sigma_y^2, \sigma_{XY}$ known

- Test name:
- Test Statistic
- Test Reference Distribution
- Critical Value/ Rejection region
- Confidence interval

field: Data setting $X_1, \ldots, X_n$ iid F_x,
$Y_1, \ldots, Y_n$ iid F_y,
X_i not independent Y_i,
$(X_1, Y_1), \ldots, (X_n, Y_n)$ iid F_{XY} $Cov(X_i, Y_i) = \sigma_{XY}$, $Cov(X_i, Y_j) = 0$.
Estimate $d = \mu_x - \mu_y$, when $\sigma_x^2, \sigma_y^2, \sigma_{XY}$ known

- Test name: Paired z-test
- Test Statistic: $z(d_0) = \frac{(\bar{X} - \bar{Y}) - d_0}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_Y^2}{n} - 2\frac{\sigma_{XY}}{n}}} = \frac{\bar{D} - d_0}{\sqrt{\frac{\sigma_D^2}{n}}}$
- Test Reference Distribution: Under H_0, $z(d_0)$ aprox $\sim N(0, 1)$
- Critical Value/ Rejection region: Same as normal
- Confidence interval :

$$(\bar{X} - \bar{Y}) \pm z(1 - \frac{\alpha}{2})\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_Y^2}{n} - 2\frac{\sigma_{XY}}{n}} = \bar{D} \pm z(1 - \alpha/2)\sqrt{\frac{\sigma_D^2}{n}}$$

NOTE:

field: 100042

field: Data setting $X_1, \ldots, X_n$ iid F_x, $Y_1, \ldots, Y_n$ iid F_y
X_i not independent from Y_i
$(X_1, Y_1), \ldots, (X_n, Y_n)$ iid F_{XY} $Cov(X_i, Y_i) = \sigma_{XY}$, $Cov(X_i, Y_j) = 0$
Estimate $d = \mu_x - \mu_y$, when $\sigma_x^2, \sigma_y^2, \sigma_{XY}$ unknown

- Test name:
- Estimate of σ_{XY}
- Estimate of $Var(\bar{X} - \bar{Y})$

field: Data setting $X_1, \ldots, X_n$ iid F_x, $Y_1, \ldots, Y_n$ iid F_y, X_i not independent Y_i, $(X_1, Y_1), \ldots, (X_n, Y_n)$ iid F_{XY} $Cov(X_i, Y_i) = \sigma_{XY}$, $Cov(X_i, Y_j) = 0$
Estimate $d = \mu_x - \mu_y$, $\sigma_x^2, \sigma_y^2, \sigma_{XY}$ unknown

- Test name: Paired Data t-test
- Estimate of $\sigma_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$
- Estimate of $Var(\bar{X} - \bar{Y}) = \frac{s_d^2}{n} = \frac{s_x^2}{n} + \frac{s_Y^2}{n} - 2\frac{s_{XY}}{n}$
- Test Statistic: $t(d_0) = \frac{(\bar{X}-\bar{Y})-d_0}{\sqrt{\frac{s_x^2}{n}+\frac{s_Y^2}{n}-2\frac{s_{XY}}{n}}} = \frac{\bar{D}-d_0}{\sqrt{\frac{s_D^2}{n}}}$
- Test Reference Distribution: If differences are Normal (note X,Y Normal does not imply Differences are normal unless X,Y are jointly multivariate-normal) Under H_0, $t(d_0) \sim t_{n-1}$ (exact distribution)
- Critical Value/ Rejection region Same as t
- Confidence interval

$$(\bar{X} - \bar{Y}) \pm t_{n-1,(1-\frac{\alpha}{2})} \sqrt{\frac{s_x^2}{n} + \frac{s_Y^2}{n} - 2\frac{s_{XY}}{n}} = \bar{D} \pm t_{n-1,(1-\frac{\alpha}{2})} \sqrt{\frac{s_d^2}{n}}$$

- Equivalent to a one sample - t-test on the differences

NOTE:

field: 100043

field: Data setting $X_1, \dots, X_m$ iid Bernoulli(p_x), $Y_1, \dots, Y_n$ iid Bernoulli(p_y), Test $H_0 : p_x - p_y = 0$

- Test name:
- Test Statistic
- Test Reference Distribution
- Critical Value/ Rejection region
- Confidence interval

field: Data setting $X_1, \dots, X_m$ iid Bernoulli(p_x), $Y_1, \dots, Y_n$ iid Bernoulli(p_y), Test $H_0 : p_x - p_y = 0$

- Test name: Binomial proportions two-sample z-test
- Test Statistic:
$$z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\hat{p}_c(1 - \hat{p}_c(\frac{1}{m} + \frac{1}{n}))}}$$
Where $\hat{p_c} = \frac{m\hat{p}_x + n\hat{p}_y}{m+n} = \frac{b+d}{N}$
- Test Reference Distribution: Under $H_0 : z$ approx $\sim N(0,1)$
- Critical Value/ Rejection region: Same as regular 2-sample
- Confidence interval:
$$\hat{p}_x - \hat{p}_y \pm z_{1-\alpha/2}\sqrt{(\frac{\hat{p}_x(1-\hat{p}_x)}{m} + \frac{\hat{p}_y(1-\hat{p}_y)}{n})}$$

NOTE:

field: 100044

field: Multinomial sampling

field: Collection of random samples, recording what group they are in: Can estimate $P(X = x|G = g)$, where G is the group

NOTE:

field: 100045

field: Two-Sample Binomial sampling

field: Sample m units from group 1 and n units from group 2

NOTE:

field: 100046

field: Can we estimate $P(X = x|G = g)$ with binomial sampling

field: Cannot estimate

NOTE:

field: 100047

field: Can we estimate $P(X = x|G = g)$ with multinomial sampling?

field: Yes

NOTE:

field: 100048

field: $E(g(T)) =$

field: $E(g(T)) \neq g(E(T))$

NOTE:

field: 100049

field: Reason for performing transformations on data

field: Some tests are FSE only when population distribution is Normal (otherwise the methods are asymptotically exact), requiring a large n. Transformations that improve approximation of normality make Normal-based methods perform more exactly

NOTE:

field: 100050

field: Data setting $X_1, \ldots, X_m$ iid Bernoulli(p_x), $Y_1, \ldots, Y_n$ iid Bernoulli(p_y), Test $H_0 : p_x - p_y = 0$ (Association/independent/relationship)

- Test name:
- Test Statistic
- Test Reference Distribution
- Critical Value/ Rejection region

field: Data setting $X_1, \ldots, X_m$ iid Bernoulli(p_x), $Y_1, \ldots, Y_n$ iid Bernoulli(p_y), Test $H_0 : p_x - p_y = 0$ (Association/independent/relationship)

- Test name: Pearson's Chi-squared Test
- Test Statistic: $X = \sum_{i,j \in \{1,2\}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ Where $O_{ij} = n_{ij}$ and $E_{ij} = \frac{R_i C_j}{N}$
- Test Reference Distribution: Under H_0 $X \sim \chi^2_1$
- Critical Value/ Rejection region: Reject for $X > \chi^2_1(1 - \alpha)$
- Note: Equal to to sided z-test for binomial proportions: $X = z^2$

NOTE:

field: 100051

field: Data setting $X_1, \ldots, X_m$ iid Bernoulli(p_x)
$Y_1, \ldots, Y_n$ iid Bernoulli(p_y)
Test $H_0 : p_x = p_y$ (Fisher)

field:

- Test for (No association between response variable X and grouping variable G)
- Test name: Fisher's Exact Test (of homogeneity of proportions)
- Test Statistic: Probability of observed table conditioning on margins: Compute all tables with the same margin totals: $\frac{\binom{C_1}{O_{11}}\binom{C_2}{O_{12}}}{\binom{N}{R_1}}$
- pvalue: Sum of probability of all tables more extreme than observed table More Extreme:
  - $p_x > p_y$ More extreme = larger O_{12}
  - $p_x < p_y$ More extreme = smaller O_{12}
  - $p_x \neq p_y$ More extreme = less likely table

Note, requires a table that keeps track of the pairs

Measurement 2	Measurement 1 No	Yes	Total
No	a	b	R_1
Yes	c	d	R_2
Total	C_1	C_2	n

NOTE:

field: 100052

field: Data setting $X_1, \ldots, X_m$ iid Bernoulli(p_x), $Y_1, \ldots, Y_n$ iid Bernoulli(p_y), Test $H_0 : p_x = p_y$ Binomial sampling scheme

field:

- Test name: Log Odds - test $H_0 : \omega = 1$
- Test Statistic: $\hat{\omega} = \frac{ad}{bc}$, $z = \frac{\log(\hat{\omega})}{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}$
- Test Reference Distribution $\log(\hat{\omega})$ approx $\sim N(\log(\omega), \frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d})$, z approx $\sim N(0,1)$
- Critical Value/ Rejection region
- Confidence interval $(\hat{\omega}e^{-z(1-\frac{\alpha}{2})\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}, \hat{\omega}e^{z(1-\frac{\alpha}{2})\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}})$
- : $\omega > 1, p_1 > p_2$, $\omega = 1, p_1 = p_2$, small p_1, p_2, $\omega = p_1/p_2$ = relative risk

Note, requires a table that keeps track of the pairs

Measurement 2	Measurement 1 No	Yes	Total
No	a	b	R_1
Yes	c	d	R_2
Total	C_1	C_2	n

NOTE:

field: 100053

field: Data setting $X_1, \ldots, X_n$ iid Bernoulli(p_x), $Y_1, \ldots, Y_n$ iid Bernoulli(p_y)
X, Y not independent (paired)
Test proportions equal in groups (equally likely/probability)

field:

- Note, requires a table that keeps track of the pairs

Measurement 2	Measurement 1 No	Yes	T
No	a	b	
Yes	c	d	
Total	C_1	C_2	

- Test name: McNemar's Test
- Test Statistic: $z = \frac{b-c}{\sqrt{b+c}}$
- Test Reference Distribution: $z \sim N(0,1)$, $z^2 \sim \chi^2_1$
- Critical Value/ Rejection region: Two sided rejecct $|z| > z_{(1-\alpha/2)}$
- Note equivalent to performing a paired t-test on the differences:

$$t = \frac{b-c}{\sqrt{\frac{n}{n-1}(b+c-\frac{(b-c)^2}{40})}}$$

  compare to t_{n-1}

NOTE:

field: 100054

field: Data setting: n observations, record Group 1 and Group 2, where each group takes on > 2 values
Test if there is an association between the groups

field:

- Test name: Pearsons χ^2
- Test Statistic: $X = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$, where $E_{ij} = \frac{n_i n_j}{N}$
- Test Reference Distribution: Under H_0, X approx $\sim \chi^2_{(r-1)(c-1)}$
- Note not FSE, but performance is good if $E_{ij} > 5$

NOTE:

field: 100055

field: Data setting $X_1, \ldots, X_m$ iid F_x, $Y_1, \ldots, Y_n$ iid F_y
Test $m_x = m_y$ (with additive assumption)

- Assumptions
- Process
- Ties
- Continuity correction
- Consistency

field:

- Test name: Wilcoxon Rank-Sum (Mann-Whitney U-test)
- Note this is only a test of medians only if just additive effect - F_x is just a shift from F_y (shape and scale must be same) (but then just the same as a test of mean, 10th percentile, min, $F_x = F_y$ etc)
- If No additive assumption - test of $H_0 : P(X > Y) = .5$
- Process:
  - Combine samples
  - Rank the observations in combined sample from smallest to largest (1 to $n + m$)
  - Add ranks of the smaller group (assume wlog that X is the smaller group)
- pvalue: Calculate using permutations: Count number of permutations that lead to a R value more extreme than observed out of total permutations ($\binom{n+m}{m}$)
- Test Statistic: R sum of the ranks, or $z = \frac{R - \frac{m(m+n+1)}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}$
- Test Reference Distribution: If there was no difference between two populations, then each rank has equal chance of being assigned to group 1 (belongs to X: $p = \frac{m}{n+m}$) Normal approximation: $R \dot{\sim} N(\frac{m(m+n+1)}{2}, \frac{mn(m+n+1)}{12})$, $z \dot{\sim} N(0, 1)$

- Notes: If ties, assign ranks, and then average ranks of tied values
- Continuity correction to normal distribution: add .5 to R if lower probability, subtract .5 from R if upper probability (ie 1 - pnorm())
- Not consistent test unless under additive assumption. IS consistent test of $H_0 : P(X > Y) = .5$

NOTE:

field: 100056

field: Data setting $X_1, \ldots, X_m$ iid F_x, $Y_1, \ldots, Y_n$ iid F_y
Test $m_x = m_y$ (medians)

field:

- Test name: Mood's Test for Equality of Population Medians
- Process:
  - Find combined sample median $\hat{m}$
  - Calculate $\hat{p}_x$ = proportion of Xs greater than $\hat{m}$, $\hat{p}_y$, proportion of Ys greater than $\hat{m}$
  - Conduct two sample binomial z-test(Pearsons chi-squared test) or Fisher's exact test
  - Test statistic:
$$z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\hat{p}_c(1 - \hat{p}_c(\frac{1}{m} + \frac{1}{n}))}}$$
Where $\hat{p_c} = \frac{m\hat{p}_x + n\hat{p}_y}{m+n} = \frac{b+d}{N}$

NOTE:

field: 100057

field: Data setting $X_1, \ldots, X_n$ iid F_x, $Y_1, \ldots, Y_n$ iid F_y
Test some statistic W

- Test name:
- Process

field:

- Test name: Permutation test
- Process: Permute group labels across observations and recalculate statistic for each permutation to create permutation distribution - calculate p-values using the permutation distribution
- Performance: Many settings (like medians equal), will not reject correctly (even in large samples) if the medians are equal, but the distributions differ
- Permutation hypothesis is that the observations from the two pouplations are exchangable (ie same population distributions, not just equal medians)

NOTE:

field: 100058

field: Data setting: Estimate value of nuisance parameter

field:

- Test name: Bootstrap
- Process: Since the empirical distribution function converges to the true distribution function, we can use samples from the empirical distribution to approximate how samples from the true distribution would behave.
- Confidence interval: $100(\alpha/2)$ largest resampled statistic $100(1-(\alpha/2))$ largest resampled statistic

NOTE:

field: 100059

field: Data setting: $X_1, \ldots, X_m$ iid N, $Y_1, \ldots, Y_n$ iid $N(\mu, \sigma^2)$.
$H_0 : \sigma_x^2 = \sigma_y^2$ vs $H_1 : \sigma_x^2/\sigma_y^2 = r \neq 1$

field:

- Test name: F
- Recall $s_x^2 = \frac{1}{n-1} \sum_{i=1}^m (X_i - \bar{X})^2$
- Note that $\frac{(m-1)s_x^2}{\sigma_x^2} \sim \chi^2_{m-1}$, $\frac{(n-1)s_y^2}{\sigma_y^2} \sim \chi^2_{n-1}$,
- Test Statistic: $F(r) = \frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2} = \frac{s_x^2}{s_y^2}\frac{1}{r}$
- Test Reference Distribution: Under $H_0 : F(r) \sim F_{m-1,n-1}$
- Critical Value/ Rejection region
  - $\sigma_x^2/\sigma_y^2 > r$ Reject for $F(r) > F_{m-1,n-1}(1-\alpha)$
  - $\sigma_x^2/\sigma_y^2 > r$ Reject for $F(r) > F_{m-1,n-1}(\alpha)$
  - $\sigma_x^2/\sigma_y^2 \neq r$ Reject for $F(r) > F_{m-1,n-1}(1-\alpha/2)$ or $F(r) < F_{m-1,n-1}(\alpha/2)$
- Performance: Not Well if underlying population is not normal: Not FSE or AE (but is consistent) - don't use if population is not normal

NOTE:

field: 100060

field: Data setting: $X_1, \ldots, X_m$ iid F_x, $Y_1, \ldots, Y_n$ iid F_y. $H_0 : \sigma_x^2 = \sigma_y^2$

- Test name:
- Process:
- Interpretation
- Assumptions

field:

- Test name: Levene's Test
- Process:
  - Construct new variables:
    * $U_i = |X_i - med(X)|$ or $(X_i - med(X))^2$ or $|X_i - \bar{X}|$ or $(X_i - \bar{X})^2$
    * $V_i = |Y_i - med(Y)|$ or $(Y_i - med(Y))^2$ or $|Y_i - \bar{Y}|$ or $(Y_i - \bar{Y})^2$
  - Perform two-sample t test on U_i and V_i (use Welch)
- Interpretation: If last option used, can be a test in difference in population variances
- Assumptions:
  - Independence
  - Large sample sizes, so t-test assumptions are met
- Note: dont use as a test to determine which t-test version to use

NOTE:

field: 100061

field: Data setting: Data setting $X_1, \ldots, X_m$ iid F_x, $Y_1, \ldots, Y_n$ iid F_y. Test $H_0 : F_x = F_y$

- Test name
- Test statistic

field: Data setting: Data setting $X_1, \ldots, X_m$ iid F_x, $Y_1, \ldots, Y_n$ iid F_y. Test $H_0 : F_x = F_y$

- Test name: Two-sample Kolmogorov-Smirnov Test
- Test statistic: $D = sup_x|\hat{F}_x(x) - \hat{F}_y(y)|$ ie the largest distance between the empirical CDF for X and Y

- Reject for large values of $\sqrt{\frac{mn}{m+n}}D$
- Only for continuous distributions, for discrete distributions, use Pearsons χ^2

NOTE:

field: 100062

field: Multiple 2x2 tables under k different conditions
$p_{xj} = P(X = 1 \text{ in Table } j), p_{yj} = P(Y = 1 \text{ in Table } j)$
Test $H_0 : p_{xj} = p_{yj}$ for all j

field:

- Test name: Mantel-Haenszel Test
- Test statistic: $\omega_j = \frac{p_{xj}(1-p_{xj})}{p_{yj}(1-p_{yj})}$, $H_0 : \omega_j = 1$ for all j

$$E(n_{X1j}) = \mu_{X1j} = \frac{n_{X\cdot j} n_{\cdot 1j}}{n_{\cdot j}}, V(n_{X1j}) = \sigma^2_{X1j} = \frac{n_{X\cdot j} n_{Y\cdot j} n_{\cdot 1j} n_{\cdot 0j}}{n^2_{\cdot\cdot j}(n_{\cdot\cdot j} - 1)}$$

$$C = \frac{[\sum_j (n_{X1j} - \mu_{X1j})]^2}{\sum_j \sigma^2_{X1j}}$$

- Under H_0 $C \dot{\sim} \chi^2(1)$
- Assumes the odds-ratios are the same in all k tables

NOTE:

field: 100063

field: Test for data setting:

- Sample 1: $X_{1,1}, \ldots, X_{1n_1}$ from population 1 with mean μ_1,
- Sample 2: $X_{2,1}, \ldots, X_{2n_2}$ from population 2 with mean μ_2
- $\ldots$
- Sample M: $X_{M,1}, \ldots, X_{Mn_M}$ from population M with mean μ_M

field:

- Test name: ANOVA
- Assumptions
  - Independence within and between groups
  - Populations (approximately) normal
  - Equal variances
- Estimate of common variance $s_p = \frac{(n_1-1)s_1^2+\cdots+(n_M-1)s_M^2}{(n_1-1)+\cdots+(n_M-1)}$
- Could use two-sample-t test on two population means
- Could test are population means 1 through M equal to each other?
- Compare the variability between groups to the variability withing groups
- Sum of squares within groups:

  $$SSW = (n-M)s_p^2 = \sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 + \cdots + \sum_{i=1}^{n_M}(X_{Mi} - \bar{X}_M)^2$$

  degrees of freedom: $n - M$
- Sum of squares total

  $$SST = \sum_{i=1}^{n_1}(X_{1,i} - \bar{X})^2 + \cdots + \sum_{i=1}^{n_M}(X_{M,i} - \bar{X})^2$$

  degrees of freedom: $n - 1$
- Sum of squares between groups:

  $$SSB = SST - SSW = \sum_{j=1}^{M} n_j(\bar{X}_j - \bar{X})^2$$

  df: $(n-1) - (n-M) = M - 1$
- Test statistic:

  $$F = \frac{MSB}{MSW} = \frac{SSB/(M-1)}{SSW/(n-M)}$$
- Reference distribution: Under $H_0, F \sim F_{M-1,n-M}$

tags: Methods2

NOTE:

field: 100064

field: Vectors $\mathbf{x}$ and $\mathbf{y}$ orthogonal

field: Vectors $\mathbf{x}$ and $\mathbf{y}$ orthogonal (perpendicular) if $(x, y) = \mathbf{x}^t\mathbf{y} = 0$

NOTE:

field: 100065

field: A matrix $\mathbf{A}$ is orthogonal if:

field: A matrix $\mathbf{A}$ is orthogonal if $\mathbf{A}^t\mathbf{A} = \mathbf{A}\mathbf{A}^t = \mathbf{I}_n$

NOTE:

field: 100066

field: A set of n vectors are linearly dependent

field: A set of n vectors are linearly dependent if there exist constants $c_1, \ldots c_n$ not all 0 such that $\sum_{j=1}^{n} c_j\mathbf{x}_j = 0$

NOTE:

field: 100067

field: Inverse of a square matrix: $\mathbf{A}_{n\times n}$

field: The matrix that will satisfy $AA^{-1} = I$

NOTE:

field: 100068

field: Inverse of $\mathbf{A}$, $\mathbf{A}^{-1}$ where $\mathbf{A}$ is 2×2

field: $\mathbf{A}^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

NOTE:

field: 100069

field: A square matrix is invertible if:

field: A square matrix is invertible if the columns (rows) are linearly independent. (If the columns are not independent, the matrix is called singular)

NOTE:

field: 100071

field: Square of matrix $\mathbf{A}$

field: $\mathbf{A}\mathbf{A}^t$

NOTE:

field: 100072

field: Norm of a vector $|\mathbf{x}|$

field: $|\mathbf{x}| = \sqrt{\sum_{j=1}^{p} x_j^2}$

NOTE:

field: 100073

field: Determinant of a 2×2 matrix

field: $\left| \begin{pmatrix} a & b \\ c & d \end{pmatrix} \right| = ad - bc$

NOTE:

field: 100074

field: Trace of a square matrix

field: Sum of the diagonal elements

NOTE:

field: 100075

field: Rank of a matrix

field: Number of linearly independent columns

NOTE:

field: 100076

field: Eigenvalue and eigenvector

field: λ is an eigen value and $\mathbf{u}_{n\times 2}$ is the eigen vector of $\mathbf{A}_{n\times n}$ if $\mathbf{Au} = \lambda\mathbf{u}$

- A real symmetric matrix has n eigen values and n eigen vectors, and each are orthogonal to each other
- Roots of $det(\mathbf{A} - \lambda\mathbf{I})$ determine the eigenvalues of A

NOTE:

field: 100077

field: Matrix properties

- $(AB)^t =$
- $(A+B)^t =$
- $(AB)^{-1} =$
- $(\mathbf{A}^{-1})^t =$

field: Matrix properties

- $(AB)^t = B^t A^t$
- $(A+B)^t = A^t + B^t$
- For invertible matrices $(AB)^{-1} = B^{-1}A^{-1}$
- For invertible matrices $(\mathbf{A}^{-1})^t = (\mathbf{A}^t)^{-1}$

NOTE:

field: 100079

field:

$$E(Y_i|X_{i1},\ldots,X_{ip}) =$$

Where Y_i is the ith response and X_{ij} is the ith value of the jth predictor

field: Since the error terms ϵ_i are independent and normally distributed with mean 0,

$$E(Y_i|X_{i1},\ldots,X_{ip}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

NOTE:

field: 100080

field: Matrix form of linear Model and data

field:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

NOTE:

field: 100081

field: Assumptions of a linear model

field:

- Linearity: $E(\epsilon_i) = 0$ or $E(\epsilon) = \mathbf{0}$ or $E(\mathbf{Y}) = \mathbf{X}\beta$
- Constant variance $V(Y_i) = \sigma^2 = Var(\epsilon_i)$ or $V(\epsilon) = \sigma^2\mathbf{I}_n$
- Normality Y_i follows normal distribution, equivalently, ϵ_i follows normal distribution
- Independence Y_i are indepednent equivalently under normality $Cov(\epsilon_i, \epsilon_j) = 0$

NOTE:

field: 100082

field: Interpretation of intercept of linear model

field: Mean response when all explanatory variables are 0

NOTE:

field: 100083

field: Interpretation of slopes of linear model + example when $p = 0$

field: Change in mean response for 1 unit change in the value of the explanatory, keeping all other variables constant. When $p = 2$

$$E(Y|X_1 + 1, X_2) - E(Y|X_1, X_2) = \beta_1$$

NOTE:

field: 100084

field: Reason for $g - 1$ indicator variables for a variable with g values

field: The model matrix $X_{n\times(p+1)}$ needs to be full column rank - $\mathbf{X}^t\mathbf{X}$ needs to be non-singular If there is no intercept, we can include all groups, but interpretation will be different

NOTE:

field: 100085

field: Interpretation of slope coefficient for indicator variable β

field: Difference in expected value of Y between group value a and b where a is the associated value for β_j and b is the base category

NOTE:

field: 100086

field:

- $E(\mathbf{AU} + \mathbf{b}) =$
- $V(\mathbf{AU} + \mathbf{b}) =$

field:

- $E(\mathbf{AU} + \mathbf{b}) = \mathbf{A}E(\mathbf{U}) + \mathbf{b}$
- $V(\mathbf{AU} + \mathbf{B}) = \mathbf{A}V(\mathbf{U})\mathbf{A}^t$

NOTE:

field: 100087

field: Least squares estimate of β (process to find)

field: Minimize the squared error loss ($L(\beta)$) with respect to β

$$L(\beta) = \sum_{i=1}^{n} Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}))^2 = (\mathbf{Y} - \mathbf{X}\beta)^t(\mathbf{Y} - \mathbf{X}\beta)$$

NOTE:

field: 100088

field:

$$\frac{\partial}{\partial \beta} L(\beta) =$$

(Finding estimate for least squares)

field:

$$
\begin{aligned}
\frac{\partial}{\partial\beta}L(\beta) &= \frac{\partial}{\partial\beta}(\mathbf{Y}-\mathbf{X}\beta)^t(\mathbf{Y}-\mathbf{X}\beta) \\
&= \frac{\partial}{\partial\beta}\mathbf{Y}^t\mathbf{Y} - \beta^t\mathbf{X}^t\mathbf{Y} - \mathbf{Y}^t\mathbf{X}\beta - \beta^t\mathbf{X}^t\mathbf{X}\beta \\
&= 0 - \mathbf{X}^t\mathbf{Y} - \mathbf{X}^t\mathbf{Y} + 2\mathbf{X}^t\mathbf{X}\beta \\
\mathbf{X^tX}\beta &= \mathbf{X}^t\mathbf{Y} \\
\beta &= (X^tX)^{-1}X^tY
\end{aligned}
$$

NOTE:

field: 100089

field: Least squares estimate of $\hat{\beta}$

field:

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$$

(if $\mathbf{X}^t\mathbf{X}$ is invertible)

NOTE:

field: 100090

field: Residual

field: $e_i = Y_i - \hat{Y}_i$, $\mathbf{e}_{n\times 1} = \mathbf{Y} - \hat{\mathbf{Y}}$

NOTE:

field: 100091

field: Vector of fitted values (linear regression)

field: $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$

NOTE:

field: 100092

field: Projection matrix

field: Hat matrix

$$\mathbf{H_{n\times n}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$$

H_{ij} is the rate at which the ith fitted value changes as we vary the jth observation (influence)

NOTE:

field: 100093

field: Properties of projection matrix H

field:

- $\mathbf{H}$ and $\mathbf{I}-\mathbf{H}$ are symmetric matrices
- $\mathbf{HX} = \mathbf{X}$
- $(\mathbf{I}-\mathbf{X})\mathbf{X} = \mathbf{0}$
- $\mathbf{H}^2 = \mathbf{H}$
- $(\mathbf{I}-\mathbf{H})\mathbf{H} = 0$
- $\mathbf{X}^t\mathbf{e} = 0$

NOTE:

field: 100094

field: Unbiased estimate of σ^2 for linear models

field: $\hat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} e_i^2 = \frac{1}{n-(p+1)} \mathbf{e}^t \mathbf{e}$

NOTE:

field: 100095

field: $\mathbf{e}^t \mathbf{e} =$

field: $\mathbf{e}^t \mathbf{e} = \mathbf{Y}^t \mathbf{Y} - \mathbf{Y}^t \mathbf{H} \mathbf{Y}$

NOTE:

field: 100096

field: $E(\hat{\beta}) =$

field: $E(\hat{\beta}) = E((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}) = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t E(\mathbf{Y}) = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\beta = \beta$
So $\hat{\beta}$ is an unbiased estimate

NOTE:

field: 100097

field: Gauss - Markov Theorem

field: If $E(\mathbf{Y}) = \mathbf{X}\beta$ and $V(\mathbf{Y}) = \sigma^2\mathbf{I}$, then the least squares estimate $\hat{\beta}$ has the least variance among all linear unbiased estimators of β. (BLUE) Note that non-normal (or iid) residuals is not nescessary, just must be uncorrelated.

NOTE:

field: 100098

field: $V(\hat{\beta}) =$

field: $V(\hat{\beta}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$

NOTE:

field: 100099

field: $E(\hat{\sigma}^2) =$

field: $E(\hat{\sigma}^2) = \sigma^2$

NOTE:

field: 100100

field: If $\mathbf{X}_{p\times 1}$ has a multivariate normal distribution $N(\mu_{p\times 1}, \Sigma_{p\times p})$, then $\mathbf{AX} + b \sim$

field: If $\mathbf{X}_{p\times 1}$ has a multivariate normal distribution $N(\mu_{p\times 1}, \Sigma_{p\times p})$, then $\mathbf{AX} + \mathbf{b} \sim N(\mathbf{A}\mu + \mathbf{b}, \mathbf{A}\Sigma\mathbf{A}^t)$

NOTE:

field: 100101

field: Multivariate normal properties for $\mathbf{X}_{p\times 1} \sim N(\mu_{p\times 1}, \Sigma_{p\times p})$

field:

- $Cov(X_j, X_k) = 0$ if and only if X_j, X_k are independent (two way due to multivariate normal)
- All subsets of elements of $\mathbf{X}$ have a multivarite normal distribution
- All linear combinations of the components of X are normally distributed
- $\mathbf{a}^t\mathbf{X} \sim N(\mathbf{a}^t, \mathbf{a}^t\Sigma\mathbf{a})$ for a vector a

NOTE:

field: 100102

field: Linear Hypothesis testing single parameter $H_0 : \mathbf{c}^t\beta = d$

- $E(\mathbf{c}^t\beta) =, V(\mathbf{c}^t\beta) =$
- Test statistic and distribution
- Item of setting up hypothesis test
- Rejection Region

field: For a vector $\mathbf{c}_{(p+1)\times 1}$, we have that

- $E(\mathbf{c}^t\hat{\beta}) = \mathbf{c}^t\beta$, $V(\mathbf{c}^t\hat{\beta}) = \sigma^2\mathbf{c}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{c}$
- Thus

$$\frac{\mathbf{c}^t\hat{\beta} - \mathbf{c}^t\beta}{\sigma\sqrt{\mathbf{c}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{c}}} \sim N(0,1)$$

  and under H_0

$$T = \frac{\mathbf{c}^t\hat{\beta} - d}{\sqrt{\hat{\sigma}^2\mathbf{c}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{c}}} \sim t_{n-(p+1)}$$

- Example: testing $H_0 : \beta_1 = \beta_2, \mathbf{c} = (0, 1, -1)^t, d = 0$
- Reject $H_a : c^t\beta \neq d :$
  - $|T| > t_{n-(p+1)}(1 - \alpha/2)$
  - $c^t\beta > d, T > t_{n-(p+1)}(\alpha)$
  - $c^t\beta < d : T < t(1 - \alpha)$

NOTE:

field: 100103

field: Confidence interval for a single parameter (linear regression slope estimate)

field:

$$\hat{\beta}_j \pm t_{n-(p-1)}(1-\alpha/2)\sqrt{\hat{\sigma}^2((\mathbf{X}^t\mathbf{X})^{-1})_{j+1,j+1}}$$

$$\mathbf{c}^t\beta \pm t_{n-(p-1)}(1-\alpha/2)\sqrt{\hat{\sigma}^2\mathbf{c}^t((\mathbf{X}^t\mathbf{X})^{-1})\mathbf{c}}$$

eg if we were testing $\beta_1 - \beta_2, c = (0, 1, -1)$

NOTE:

field: 100104

field: F statistic in matrix form

field:

- $\mathbf{K}$ is $p \times k$, $\mathbf{m}$ is $k \times 1$
- Testing $H_0 : \mathbf{K}^t\beta = \mathbf{m}$
- $F = \frac{\left((\mathbf{K}\hat{\beta}-\mathbf{m})^t(\mathbf{K}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{K}^{-1})(\mathbf{K}\hat{\beta}-\mathbf{m})\right)}{k\hat{\sigma}^2} \sim F_{k,n-p}$
- Eg$K = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}$, $m = 0$
- Tests $\beta_1 = 0$
- Note the $\mathbf{K}^t$ matrix is the coefficients of the system of linear equations for the the null hypothesis, and m is what they are equal to

NOTE:

field: 100105

field: Overall regression F-test

field: Tests if any predictors are related to the response

- Full model Ω: $\mathbf{Y} = \mathbf{X}\beta + \epsilon$
- Reduced model ω: a nested model with q estimated parameters
- eg: Reduced model: $\mathbf{Y} = \beta_0 + \epsilon$, $q = 1$
- $H_0 : \beta_1 = \ldots = \beta_p = 0$
- $F = \frac{(RSS_\omega - RSS_\Omega)/(p-q)}{RSS_\Omega/(n-p)}$

NOTE:

field: 100106

field: Analysis of Variance Table and calculated F stat

field:

Type	df	Sum of Squares	Mean SS
Regression	p	SS(Reg)	SS(Reg)/p
Residual	$n - p + 1$	SS(Res)	$\hat{\sigma}^2 = \text{SS(Res)}/n - p - 1$
Total	$n - 1$	SS(Total) = SS(Reg) + SS(Res)	$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

and $F = \frac{Mean(SSREG)}{Mean(SSRES)}$

NOTE:

field: 100107

field: Distribution of $\hat{\beta}$, where $\hat{\beta}$ are the estimated coefficients of linear regression.

field: $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1})$

NOTE:

field: 100108

field: RSS (in terms of Ω and ω)

field:

$$RSS_\Omega = \sum_{i=1}^{n} e_i^2$$

$$RSS_\omega = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

NOTE:

field: 100109

field: R^2

field:

$$R^2 = \frac{SS(Reg)}{SS(Tot)} = 1 - \frac{SS(Res)}{SS(Tot)}$$

- Where SS(Reg) is the regression sum of square: $\sum_i (\hat{y}_i - \bar{y})^2$ (fitted minus mean)
- SS(Tot) or TSS is the total sum of squares $\sum_i (y_i - \bar{y})^2$
- SS(Res) (or error sum of squares) SS_E or RSS is the residual sum of squares $\sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i$
- SS(Tot) = SS(Res) + SS(Reg)

NOTE:

field: 100110

field: Properties of the estimate of σ^2

field:

- $\hat{\sigma}^2 = \frac{|\mathbf{e}|^2}{n-(p+1)}$
- Under normality: $\frac{(n-(p+1))\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-(p+1)}$
- $\hat{\sigma}^2$ is independent from $\hat{\beta}$

NOTE:

field: 100111

field: Prediction Interval

field: Predicting a future response $\mathbf{x}_0^t\hat{\beta} \pm t_{n-p,(\alpha/2)}\hat{\sigma}\sqrt{1 + x_0^t(X^tX)^{-1}x_0}$ A 95% prediction interval for a response with (list values) is between and

NOTE:

field: 100112

field: Confidence interval

field: Confidence in mean response $\mathbf{x}_0^t\hat{\beta} \pm t_{n-p,(\alpha/2)}\hat{\sigma}\sqrt{x_0^t(X^tX)^{-1}x_0}$ With 95% confidence, the expected mean response

NOTE:

field: 100113

field: Residual Plot

field:

- Plot residuals against fitted values (so there is only 1 plot vs against explanatory variables)
- Verifies linearity and constant variance

NOTE:

field: 100114

field: Leverege

field:

- An obervarion has high leverage if the explanatory variable values of the observation are different from general pattern
- $h_i = H_{ii} = (X(X^tX)^{-1}X^t)_{ii}$
- High leverage $h_i > \frac{2(p+1)}{n}$

NOTE:

field: 100115

field: Standardized Residual

field: $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}$ Large if $|r_i| > 2$ - indicates outlier

NOTE:

field: 100116

field: Influential point

field: If fitted model depends highly on the value of the observation
Measure using cook's distance

$$D_i = \frac{(\hat{Y} - \hat{Y}_{(i)})^t(\hat{Y} - \hat{Y}_{(i)})}{(p+1)\hat{\sigma}^2} = \frac{1}{p+1} r_i^2 \frac{h_i}{(1-h_i)}$$

Where Y_i is the vector of fitted values when the model is fitted to the data without the ith observation
Moderate if > 1 Large if > 6

NOTE:

field: 100117

field: Multicollinearity

field:

- X^tX is close to singular
- Some columns are highly correlated
- there is a relationship between predictors
- leads to large standard errors
- Not a violation of assumptions, but leads to issues in interpretations
- Calculate using Condition number if > 30 than large , or Variance inflation factors $VIF_j = \frac{1}{1-R_j^2}$ where R_j^2 is R^2 from regression of the jth explanatory variable on all the other explanatory variables
- Not a problem for prediction
- Fix using selection of explanatory variables, generalized inverse, ridge regression

NOTE:

field: 100118

field: Ridge Regression parameter estimate

field: $\hat{\beta} = (X^tX + \lambda I)^{-1}X^tY$, where λ is chosen. Note these are biased estimators

NOTE:

field: 100119

field: Fix non-constant spread/variance

field:

- Transform response (box-cox)
- Use more complicated model (glm)

NOTE:

field: 100120

field: Fix non-linearity

field:

- Transform response
- Transform predictor
- allow for curvature: predictor squared, splines, gam
- use a non linear model

NOTE:

field: 100121

field: Fix Non-normality

field:

- Transform response
- more complicated models : glm

NOTE:

field: 100122

field: Missing data completely at random (MCAR)

field:

- Throwing out cases with missing data does not bias inferences
- There's no relationship between whether a data point is missing and any values in the data set, missing or observed.
- The missing data are just a random subset of the data.

NOTE:

field: 100123

field: Missing at random (MAR)

field:

- the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data.
- Probability of missingness depends only on available information, like the explanatory variables and the response variables present in the regression - impute missing data
- A better name would actually be Missing Conditionally at Random, because the missingness is conditional on another variable.

NOTE:

field: 100124

field: Model Selection methods

field:

- Sequential Methods: Backward/Forward (eliminate untill all values have p-value below critical value) Elimination
- Penalized Regression: Ridge and Lasso

NOTE:

field: 100125

field: AIC

field: Estimate the distance of a candidate model from the true model (small good)

$$n \log(RSS/n) + 2(p + 1)$$

NOTE:

field: 100126

field: BIC

field: Estimate the best parsimonious model, using a prior distribution on the parameters (small good)

$$n \log(RSS/n) + \log(n)(p + 1)$$

Where n is the number of observations, p is the number of predictors (not including intercept), and $RSS = \sum(Y_i - \hat{Y})^2 = \sum e_i^2$

NOTE:

field: 100127

field: Adjusted R^2

field: Adjusts for multiple parameters

$$1 - \frac{n-1}{n-p}(1 - R^2)$$

(large is good) (where p includes the intercept)
$\frac{MS(Reg)}{MS(Total)} = 1 - \frac{SS(Reg)/(n-p-1)}{SS(Tot)/(n-1)}$

NOTE:

field: 100128

field: Mallow's Cp

field:

$$RSS/\hat{\sigma^2} + 2p - n$$

(small good)

NOTE:

field: 100129

field: Box-Cox Transformation

field: Transform so model is $g(Y) = X\beta + \epsilon$ where $g(y) = \frac{y^\lambda - 1}{\lambda} if \lambda \neq 0, 0$ otherwise

tags: Methods3

NOTE:

field: 100130x

field: Components of an experiment

field: Experimental units, treatment, design (how eus are allocated to treatments)

NOTE:

field: 100130

field: Model and assumptions for CRD

field: Model and assumptions for Completely randomized design

$$y_{ij} = \mu_i + \epsilon_{ij}$$

Where

- y_{ij} is the response on the jth eu in the ith group
- μ_i is the population mean in the ith group
- ϵ_{ij} is the random error for the jth eu in the ith group
- Assume $\epsilon_{ij} \sim iidN(0, \sigma^2)$

NOTE:

field: 100131

field: Point estimate of $\hat{\mu}_i$

field: $\hat{\mu}_i = \bar{y}_i =$ mean in the ith group

NOTE:

field: 100132

field: Point estimate of $\hat{\sigma}^2$

field:

$$\begin{aligned} \hat{\sigma}^2 = MSE = \frac{\text{error sum of squares}}{df} = \frac{\text{residual SS}}{df} \quad (1) \\ = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2}{N-g} \quad (2) \\ = s^2 \quad (3) \end{aligned}$$

Where

- g is the number of groups

- N is the overall sample size
- n_i the number of eus in the ith group
- Df = sample size - number of parameters = $N - g$
- ith residual = $y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i\cdot}$

NOTE:

field: 100133

field: Hypothesis test and interval estimates for μ_i in CRD

field:

- $\hat{\mu}_i = \bar{y}_i =$ sample mean of $y_{i1}, \cdots y_{in_i} \sim$ iid $N(\mu_i, \sigma^2)$
- $\bar{y}_i \sim N(\mu_i, \frac{\sigma^2}{n_i})$
- $SE(\bar{y}_i) = \sqrt{\frac{s^2}{n_i}}$
- CI: $\bar{y}_i \pm t_{(a/2, N-g)} \sqrt{\frac{s^2}{n_i}}$
- $H_0 : \mu_i = 0$
- $t = \frac{\bar{y}_i}{\sqrt{s^2/n_i}} \sim t_{(N-g)}$

NOTE:

field: 100134

field: Cell Means Parametrization (eg $g = 3, n_i = 2$)

field:

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{32} \end{pmatrix}$$

NOTE:

field: 100135

field: Regression parametrization (eg $g = 3, n_i = 2$)

field: Code categorical variables using indicators $y_{ij} = \beta_0 + \beta_1 X_{1,ij} + \beta_2 X_{2,ij} + \epsilon_{ij}$

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{32} \end{pmatrix}$$

NOTE:

field: 100136

field: Factor (Treatment) Effects Parametrization

field:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Where

- μ = overall mean: average of μ_i

- α_i = effect of level i of the treatment factor, deviation away from μ associated with the ith treatment

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{32} \end{pmatrix}$$

Note that $\alpha_3 = -\alpha_1 - \alpha_2$

NOTE:

field: 100137

field: Extra Sum of Squares F- test

field:

- Compares full and reduced models

$$F = \frac{(SS_E(\text{red}) - SS_E(\text{full}))/(df(\text{red}) - df(\text{full}))}{SS_E(\text{full})/df(\text{full})}$$

- Can use to test for differences across the group means

$$H_0 : \mu_1 = \ldots = \mu_g = \mu$$

$$H_A : \mu_i \neq \mu_j \text{ for some } i \neq j$$

- Reduced model: $y_{ij} = \mu + \epsilon_{ij}$
- Full model: $y_{ij} = \mu_i + \epsilon_{ij}$
- $SS_E = \sum_j (y_j - \hat{y}_j)^2$ = residual SS
- $SS_E(\text{full}) = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$ Where $\bar{y}_{i\cdot}$ is the fitted value for obs in ith group

- $SS_E(\text{red}) = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2$ where $\bar{y}_{\cdot\cdot} = \sum_i \sum_j y_{ij}/N$ mean of all obs
- df(full) $= N - g$
- df(red) $= N - 1$
- SSE(red) - SSE(full) = SSTreatment for CRD
- Reduced model will have more unexplained variation

NOTE:

field: 100138

field: CRD ANOVA Table

field:

	DF	SS	MS	F
Treatment	$g-1$	SS(Trt)	SS(Trt)/$(g-1)$	MS(trt)/MS(E) $\sim F_{g-1,N-g}$
Error	$N-g$	SS(E)	SS(E)/$(N-g)$	
Total	$N-1$	SS(T)		

NOTE:

field: 100139

field: Distribution of SS(Total)/σ^2, SS(Treatment)/σ^2 and SS(E)/σ^2

field: χ^2_{N-1}, χ^2_{g-1} , χ^2_{N-g}

NOTE:

field: 100140

field: $E(MS_{\text{Trt}}) =$

field:

- If H_0 true, then $E(MS_{\text{Trt}}) = \sigma^2$
- If H_A true, then $E(MS_{\text{Trt}}) > E(MS_E)$

NOTE:

field: 100141

field: $E(MS_{\text{E}}) =$

field: $E(MS_{\text{E}}) = \sigma^2$

NOTE:

field: 100142

field: Contrast

field: A contrast is a linear combination of treatment means where the coefficients sum to 0 $C = \sum_{i=1}^{g} w_i \mu_i$ where $\sum_{i=1}^{g} w_i = 0$ Examples:

- $\frac{\mu_1+\mu_2+\mu_3}{3} - \mu_4$, $C = 1/3, 1/3, 1/3, -1$

NOTE:

field: 100143

field: Hypothesis test of contrast

field:

- $\hat{C} = \sum_{i=1}^{g} w_i \bar{y}_{i\cdot}$
- $V(\hat{C}) = V\left(\sum_{i=1}^{g} w_i \bar{y}_{i\cdot} \right) = \sum_{i=1}^{g} w_i^2 \frac{\sigma^2}{n_i}$
- $\hat{V}(\hat{C}) = \sum_{i=1}^{g} w_i^2 \frac{MS_E}{n_i} = \sum_{i=1}^{g} w_i^2 \frac{\hat{\sigma}^2}{n_i}$
- CI: $\hat{C} \pm t_{(1-\alpha/2, N-g)} SE(\hat{C})$
- $t = \frac{\hat{C}-0}{SE(\hat{C})} \sim t_{N-g}$
- Eg if $C = \mu_1 - \mu_4$ a test of $C = 0$ is testing $\mu_1 = \mu_4$

NOTE:

field: 100144

field: Contrast sums of squares

field: $SS_{\text{Contrast}} = SS_E(\text{reduced}) - SS_E(\text{full})$

- The full model is the separate means model $y_{ij} = \mu_i + \epsilon_{ij}$
- The reduced model is the full model with the restriction $H_0 : C = 0$ imposed on the μ_i
- Eg: $C = \frac{\mu_1+\mu_2+\mu_3}{3} = \mu_4$ Full model parameter vecotr $(\mu_1, \cdots, \mu_4)^t$, reduced model parameter vector: (μ_1, μ_2, μ_3) with $\mu_4 = \frac{\mu_1+\mu_2+\mu_3}{3}$
- df full $= N-4$, df reduced $= N-3$, df contrast $= 1 = (N-3)-(N-4)$

NOTE:

field: 100145

field: Orthogonal contrasts

field: Contrasts C_1 and C_2 are orthogonal if $\sum_{i=1}^{g} \frac{w_i w_i^*}{n_i} = 0$
We usually only consider orthogonal contrasts when $n_i = n$ (balanced design)
With g treatments, we can have at most $g - 1$ orthogonal contrasts
If contrasts are orthogonal $SS(trt) = SS(C1) + \ldots + SS(C_{g-1})$

NOTE:

field: 100146

field: Orthogonal polynomial contrasts and polynomial regression

field:

- When data are balanced and treatments are incremental and equally spaced, we can use orthogonal polynomial contrasts
- With g treatments, fit a $g - 1$ degree polynomial model. Fitted polynomial will fit each treatment mean exactly
- The $g-1$ degree polynomial is another parametrization of the separate means model
- The cell means model ignores the incremental nature of treatment - polynomial one doesnt
- Polynomial models imply something about interpolation
- ex: $y_{ij} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \epsilon_{ij}$ X_i is the amount of treatment in the ith group.
- SS(trt) = SS(linar) + SS(quad) + SS(cubic)

NOTE:

field: 100147

field: Design matrix for orthogonal polynomial contrasts

field:

$$X = \begin{pmatrix} 1 & 0 & 0^2 & 0^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 50 & 50^2 & 50^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 100 & 100^2 & 100^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 150 & 150^2 & 150^3 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

$\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^t$

NOTE:

field: 100148

field: Per comparison error rate

field:

- P(reject H_{0i}) when H_{0i} is true
- Usual α
- No correction for multiple comparisons

NOTE:

field: 100149

field: Experimentwise error rate

field: $\alpha_E =$ P(reject at least one H_{0i}) when H_0 is true (all H_{0i} true)

NOTE:

field: 100150

field: False Discovery rate (FDR)

field: $FDR = \frac{\text{number false rejections}}{totalnumberrejections}$, or 0 when no rejections
Allows more incorrect rejections as the number of true rejections increases

NOTE:

field: 100151

field: Strong familywise error rate

field: $\alpha_F = P(\text{at least one false rejection}) = P(FDR > 0)$

NOTE:

field: 100152

field: Tradeoff of multiple comparisons

field: Stronger error control - less powerful test

NOTE:

field: 100153

field: Bonferroni correction

field:

- K comparisons
- Fix $\alpha_F = P(\text{at least one false rejection})$ and set per comparison error rate $\alpha = \alpha_F/K$
- Reject H_{0i} if its p value is less than α_F/K
- Very strict, but easy test

NOTE:

field: 100154

field: Holm multiple comparison

field:

- K comparisons
- Sort individual p-values from small to large $p_1, \ldots, p_k$
- Reject H_{0i} if $p_i < \frac{\alpha_F}{K-i+1}$
- Note $\frac{\alpha_F}{K-i+1} \geq \frac{\alpha_F}{K}$, so Holm is more powerful than Bonferroni , but still conservative

NOTE:

field: 100155

field: Multiple comparison method: FDR

field:

- K comparisons
- Sort p-values
- Reject H_{0i} if $p_i < \frac{i \cdot FDR}{K}$
- Controls the false discovery rate

NOTE:

field: 100156

field: Scheffes method

field:

- Only method that controls α_F if we've snooped the data
- Tests all possible contrasts (all are 0)
- very conservative
- Reject $H_{0i} : C_i = 0$ if

$$\frac{SS_{C_i}(g-1)}{MS_E} > F_{\alpha_F, g-1, N-g}$$

- Confidence interval:

$$\hat{C}_i \pm \sqrt{(g-1)F_{\alpha_F, g-1, N-g}} SE(\hat{C}_i)$$

NOTE:

field: 100157x

field: Multiple comparison for pairwise comparisons

field:

- Contrasts of the form $\mu_i - \mu_j$
- For g treatment groups there are $\binom{g}{2}$ possible pairwise comparisons
- Tukey's Honestly Significant Difference (balanced)
- Tukey-Kramer (not balanced)

NOTE:

field: 100157

field: Tukey's Honestly Significant Difference (HSD)

field:

- Pairwise comparisons
- Simultaneous tests and CIs of all $C = \mu_i - \mu_j$
- Controls α_F
- CI:

$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm q_{\alpha_F, g, N-g} \sqrt{\frac{MS_E}{n}}$$

- Assumes n observations in each group (balanced)
- Where q is the studentized range distribution - dividing a statistic by the estimate of its standard error

NOTE:

field: 100158

field: Tukey-Kramer

field:

- Pairwise comparison
- If the data are not balanced (but close)
- Replace $\sqrt{\frac{MS_E}{n}}$ with $\sqrt{MS_E \frac{n_i+n_j}{2n_i n_j}}$ in
- CI:

$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm q_{\alpha_F, g, N-g} \sqrt{\frac{MS_E}{n}}$$

NOTE:

field: 100159

field: Ryan-Einot-Gabriel-Welsch Range (REGWR) test

field:

- Controls α_F
- Stepdown procedure
- Order sample means from small to large
- Test ranges, starting with largest range $\mu_{(1)} = \mu_{(g)}$
- If fail to reject, stop, conclude that no means differ. Otherwise stop down and test next largest ranges. etc

NOTE:

field: 100160

field: Dunnett's Procedure

field:

- Compare all treatments to control
-

NOTE:

field: 100161

field: Multiple Comparisons with the best MCB

field:

- Identifies either $max(\mu_i)$ or $min(\mu_i)$
- Intervals either contain 0 (not different from best) or have 0 as an endpoint, which implies they are different from the best.
- Usually done with ANOVA

NOTE:

field: 100162

field: Difference between Type I and Type III Sum of Squares

field:

- Type I is a nested model - variables are added
- Type III removes one variable

NOTE:

field: 100163

field: Effect of non-normality ("Robustness")

field:

- If tails are too long (compared to normal) estimate of variance will be too large, inference will be conservative (CI to wide, p-values too big, type I error smaller than α, lower power)
- If tails are too short, reverse is true

NOTE:

field: 100164

field: Equal variance diagnostics

field:

- Levene's test
- Plot residuals vs fitted values

NOTE:

field: 100165

field: Effect of non-constant variance + Remedy

field:

- If data are balance and variances are not too unequal, standard procedures work pretty well
- If data are unbalanced and large n_i corresponds to larger variances, procedures too conservative
- Small n_i correspond to large variances, opposite
- Remedy using Welch's ANOVA/weighted least squares, larger balanced sample

NOTE:

field: 100166

field: RF Plot

field:

- Residual-Fit Spread plot
- Left plot has sorted centered fits $\hat{y}_{ij} - \hat{\bar{y}}_{ij}$
- Right plot has sorted residuals $y_{ij} - \hat{y}_{ij}$
- Left plot shows variability explained by the model
- Right plot shows unexplained variability
- Want spread of left plot to be larger than right plot - indicates we have a good model

NOTE:

field: 100167

field: Sample size to perform a 2 sample z test

field: For a 2 sample z test $H_0 : \mu_1 = \mu_2$ with σ^2 known

$$n \geq 2(z_{\alpha/2} + z_\beta)^2 \frac{\sigma^2}{\delta^2}$$

- n sample size in each group
- $\alpha =$ Type I error rate
- $\beta =$ Type II error rate = 1-Power
- $z_{\alpha/2} =$ standard normal $1 - \alpha/2$ quantile
- $z_\beta =$ standard normal $1 - \beta$ quantile
- $\sigma^2 =$ common variance
- $\delta = \mu_1 - \mu_2$

NOTE:

field: 100168

field: Sample size for one-way ANOVA

field: Depends on the distribution when H_A is the case - non central F distribution - to find sample size, simulate repeated sampling under H_A to calculate power for different N

NOTE:

field: 100169

field: 2×2 Factorial design difference from ANOVA

field: ANOVA fits a model like, for group 1 with treatments C,F and group2 treatments HL

CH	CL	FH	FL

(ignores the structure of treatments)

Factorial design:

		Liquid L	H
Screen	C		
	F		

Uses contrasts

NOTE:

field: 100170

field: Interaction plot

field:

- If the interaction contrast is 0, then the lines will be parallel
- If we see non parallel lines, it indicates there is an interaction
- Parallel lines associated with large p values of interaction term

NOTE:

field: 100171

field: Model for a 2×2 factorial design

field: y_{ijk} is response from the kth replicate with ith level of factor A, and jth level of factor B
eg:

		B	
		$j=1$	$j=2$
A	$i=1$	y_{11k}	y_{12k}
	$i=2$	y_{21k}	y_{22k}

NOTE:

field: 100172

field: Cell means parametrization for 2×2 factorial design

field:

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

y_{ijk} is response from the kth replicate with ith level of factor A, and jth level of factor B
$\epsilon_{ijk} \sim N(0, \sigma^2)$

NOTE:

field: 100173

field: Factor effects parametrization for 2×2 design

field:

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

Where

- μ is the overall mean
- α_i effect of ith level of factor A
- β_j effect of jth level of factor B

- $(\alpha\beta)_{ij}$ interaction of ith level of A and jth level of B
- Where $\epsilon_{ijk} \sim N(0, \sigma^2)$
- $0 = \sum_{i=1}^{2} \alpha_i = \sum_{j=1}^{2} \beta_j = \sum_{i=1}^{2} (\alpha\beta)_{ij} = \sum_{i=1}^{2} (\alpha\beta)_{ij}$

NOTE:

field: 100174

field: Equivalence of cell means and factor effects parametrizations

field:

	$j = 1$	$j = 2$
$i = 1$	$\mu_{11} = \mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11}$	$\mu_{12} = \mu + \alpha_1 + \beta_2 + (\alpha\beta)_{12} = \mu + \alpha$
$i = 2$	$\mu_{21} = \mu + \alpha_2 + \beta_1 + (\alpha\beta)_{21} = \mu - \alpha_1 + \beta_1 - (\alpha\beta)_{11}$	$\mu_{22} = \mu + \alpha_2 + \beta_2 + (\alpha\beta)_{22} = \mu - \alpha$

NOTE:

field: 100175

field: Design matrix for 2×2 factorial design, where each group has 2 options

field:

$$\begin{pmatrix} y_{111} \\ \vdots \\ y_{11n} \\ y_{121} \\ \vdots \\ y_{12n} \\ y_{211} \\ \vdots \\ y_{21n} \\ y_{221} \\ \vdots \\ \\ y_{22n} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & -1 & -1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & 1 & -1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & -1 & 1 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ (\alpha\beta)_{11} \end{pmatrix} + \epsilon$$

NOTE:

field: 100176

field: General model for a 2-factor design

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

- A has levels $1 \cdots a$
- B has levels $1 \cdots b$
- $\epsilon_{ijk} \sim N(0, \sigma^2)$
- $0 = \sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{i=1}^{a} (\alpha\beta)_{ij} = \sum_{j=1}^{b} (\alpha\beta)_{ij}$
- There are $a \times b$ parameters to estimate
- $(a-1)\,\alpha_i$s , $(b-1)\,\beta_j$s, $(a-1)(b-1)\,(\alpha\beta)_{ij}$s $= ab$ total parameters

NOTE:

field: 100177

field: Parameter estimates for 2×2 factorial design

field:

- $\hat{\mu} = \bar{y}_{...}$
- $\hat{\alpha}_i = \hat{\mu}_{i.} - \hat{\mu} = \bar{y}_{i..} - \bar{y}_{...}$
- $\hat{\beta}_j = \hat{\mu}_{.j} - \hat{\mu} = \bar{y}_{.j.} - \bar{y}_{...}$
- $(\hat{\alpha\beta})_{ij} = \hat{\mu}_{ij} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\mu}$
- Where :
- $\mu_{i.}$ population mean for ith level of factor A
- $\mu_{.j}$ population mean for jth level of factor B
- α_i deviation from the overall mean associated with ith level of factor A
- $(\alpha\beta)_{ij}$ deviation of cell mean from the row column and overall mean

NOTE:

field: 100178

field: ANOVA for 2 factor design - Hypothesis test interpretation

field: Degrees of freedom:

- $A : a - 1$
- $B : b - 1$
- $AB : (a - 1)(b - 1)$
- Error $N - ab$

- Total $N-1$

Each row in anova sum of squares table gives the F value for if that row was zero, ie test all $\alpha_i = 0$ indicates that that factor has no effect

NOTE:

field: 100179

field: General factorial design (eg $8 \times 2 \times 2$)

field:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl}$$

Where

- y_{ijkl} response for lth replicate at ith level of A, jth level of B and kth level of C
- μ overall mean
- α_i effect of ith level of A
- $(\alpha\beta)_{ij}$ interaction of A and B
- $(\alpha\beta\gamma)$ interaction of A and B and C
- $\epsilon_{ijkl} iid \sim N(0, \sigma^2)$

NOTE:

field: 100180

field: Type I Type II Type III Sum of squares (unbalanced)

field:

- When data are unbalanced Type I and Type III SS are different
- I is sequential
- II is partial
- III is hierarchical

Type	SS	Effects in full model	Effects in reduced model
I	A	A	intercept only
III	A	A,B,C,AB,AC,BC,ABC	B,C,AB,AC,BC,ABC
II	A	A,B,C,BC	B,C,BC
III	AB	A,B,C,AB,AC,BC,ABC	A,B,C,AC,BC,ABC
II	AB	A,B,C,AB,AC,BC	A,B,C,AC,BC

NOTE:

field: 100181

field: Issues with Unbalanced Data for overall mean estimate and sum of squares

field:

- The fitted value for y_{ijk} is still the observed cell mean with unbalanced data
- Estimate of overall mean is not the average of all y values $\hat{\mu} \neq \bar{y}_{\cdots}$
- Issues with row and cell means
- $\bar{y}_{1\cdot\cdot} \neq \frac{\bar{y}_{11}+\bar{y}_{12}}{}$
- Type I still sums to Model Sum of Squares, but Type II and III does not (but it does for balanced data)
- Type II useful for model building
- Type III SS useful for hypothesis testing

NOTE:

field: 100182

field: Full and Reduced Type II Sum of squares, with factors A, B, C

field:

- Reduced model for facor A is largest model not containing A in any terms (ie remove interactions with a), full model adds A, but not interactions of A
- Sum of Squares for A
- Full: A,B,C,BC
- Reduced: B,C,BC
- Sum of Squares for AB
- Full: A,B,C,AB,AC,BC
- Reduced: A,B,C,AC,BC

NOTE:

field: 100183

field: Type I sum of squares, full and reduced model, with factors A,B,C

field:

- Note we could get different values depending on the order of the factors in the specified model (even p-values)
- Sequential
- Full model AB: A,B,C,AB
- Reduced model: A,B,C

- Full model: A:, A
- Reduced model: intercept only

NOTE:

field: 100184

field: Predicted values for different type of SS for factor A

field:

	Reduced model	Full model
Type I	$\hat{y}_{ijk} = \hat{\mu}$	$\hat{y}_{ijk} = \hat{\mu} + \hat{\alpha}_i$
Type II	$\hat{y}_{ijk} = \hat{\mu} + \hat{\beta}_j$	$\hat{y}_{ijk} = \hat{mu} + \hat{\alpha}_i + \hat{\beta}_j$
Type II	$\hat{y}_{ijk} = \hat{\mu} + \hat{\beta}_j + (\hat{\alpha\beta})_{ij}$	$\hat{y}_{ijk} = \hat{\mu} + \hat{\alpha}_i\hat{\beta}_j + (\hat{\alpha\beta})_{ij}$

NOTE:

field: 100185

field: Estimates for 2-factor means and interactions

field:

$$\begin{aligned}
\hat{\mu} &= \bar{y}_{\cdots} \\
\hat{\alpha}_i &= \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots} \quad \text{group mean - overall mean} \\
\hat{\beta}_j &= \hat{y}_{\cdot j\cdot} - \bar{y}_{\cdots} \\
(\hat{\alpha\beta})_{ij} &= \bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots} \quad \text{cell mean - row and col means + overall mean}
\end{aligned}$$

NOTE:

field: 100186

field: Contrasts, balanced data and orthogonal

field:

- When data are balanced, a contrast for one main effect or interaction is orthogonal to a contrast for any other main effect
- Because of orthogonality, we can estimate effects and compute SS one term at a time, and the results for that term dont depend on what other terms are in the model.
- With unbalanced data, we dont have orthogonality

NOTE:

field: 100187

field: Example of converting a categorical variable into numeric

field:

- We should know at design stage if we want to treat variable as categorical or numeric
- Linear model will be less complicated than categorical - lower model degrees of freedom

NOTE:

field: 100188

field: Missing cell

field:

- Factorial structure is missing
- Can analyze using cell means model and look at contrasts

NOTE:

field: 100189

field: When to use random effects

field:

- Levels of a factor are sampled from a larger population
- Repeating experiment would use different factor levels (ie if the choice of levels are drawn using a random sample from larger popluation)
- Need to model dependence among observations from the same level of a factor
- key word - batch
- Not all observations independent (ie boxes from same machine are similar)

NOTE:

field: 100190

field: Random effects Dependence among observations in a single group-covariance

field:

$$\begin{aligned} Cov(y_{11}, y_{12}) &= Cov(\mu + \alpha_1 + \epsilon_{11}, \mu + \alpha_1 + \epsilon_{12}) \\ &= Cov(\alpha_1, \alpha_1) + Cov(\alpha_1, \epsilon_{12}) + Cov(\epsilon_{11}, \epsilon_{12}) \\ &= \sigma_\alpha^2 \end{aligned}$$

- Last three terms are 0 because independence assumptions within and between
- α_1 and ϵ_{ij} are random, μ is fixed
- Note this model assumes positive covariance
- Note that covariance between observations in different groups is 0

NOTE:

field: 100191

field: Random Effects Model and Assumptions (1 factors)

field:

$$y_{ij} + \alpha_i + \epsilon_{ij}$$

Where

- y_{ij} = strength of jth box made by ith machine
- μ = overall mean
- α_i = effect of ith machine (allows boxes made by two different machines to systematically differ)
- ϵ_{ij} = random error

Assumptions:

- $\epsilon_{ij} \sim iidN(0, \sigma^2)$
- $\alpha_i \sim iidN(0, \sigma_\alpha^2)$
- ϵ independent from α
- Different condition from fixed effects $\sum_i \alpha = 0$

NOTE:

field: 100192

field: Estimates for μ for random effects model (1 factor) to make inference

field:

$$\hat{m}u = \bar{y}_{\cdot\cdot}$$
$$E(\hat{\mu}) = \mu$$
$$Var(\hat{\mu}) = \frac{n\sigma_\alpha^2 + \sigma^2}{N}$$
$$\hat{\mu} \sim N(\mu, \frac{n\sigma_\alpha^2 + \sigma^2}{N})$$

NOTE:

field: 100193

field: Random effects model - how to test differences among levels of the factor

field:

- If α_i were fixed, test $H_0 : \alpha_i = 0$
- If random effect, cant use this H_0 since the hypotheses must be about the parameters, and α_i are random variables
- Instead test $H_0 : \sigma_\alpha^2 = 0$ - no machine effect
- Estimated $\sigma^2, \sigma_\alpha^2$ are called variance components

NOTE:

field: 100194

field: Anova for one random factor design (A)

field:

Source	df	SS	EMS
A	$\alpha - 1$	$\sum_i \sum_j (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$\sigma^2 + n\sigma_\alpha^2$
Error	$N - \alpha$	$\sum_i \sum_j (\bar{y_{ij}} - \bar{y}_{i\cdot})^2$	σ^2
Total	$N - 1$	$\sum_i \sum_j (\bar{y}_{ij} - \bar{y}_{\cdot\cdot})^2$	

$$\frac{SS_A/(\alpha - 1)}{SS_E/(N - \alpha)} \sim F_{\alpha-1,N-\alpha}$$

Where α is the number of factors in A

NOTE:

field: 100195

field: Expected Mean Squares in one random factor ANOVA

field:

- $H_0 = \alpha_i$ is true, then $E(MS_{trt}) = \sigma^2 = E(MS_E)$
- $H_0 = \alpha_i$ is false, then $E(MS_{trt}) > E(MS_E)$
- F statistic is $\frac{MS_{trt}}{MS_E}$, and we reject H_0 if the F statistic is large
- $E(MS_A) = \sigma^2 + n\sigma_\alpha^2$. If $H_0 : \sigma_\alpha^2 = 0$ true, $E(MS_A) = \sigma^2$
- Expected mean squares tell us how to form the F statistic,
- The denominator is the MS whose expectation is equal to the numerator $E(MS)$ under H_0

NOTE:

field: 100196

field: MS_{trt} in factor design

field: MS_{trt} is MS_A for the A treatment. or MS_B if testing B treatment so $F = \frac{MS_{trt}}{MS_E}$

NOTE:

field: 100197

field: Two random factors Model and assumptions

field:

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

Where

- y_{ijk} = strength of kth box from the ith machine made by jth operator
- μ = overall mean (fixed)
- α_i = effect of ith machine
- β_j = effect of jth operator
- $(\alpha\beta)_{ij}$ = interaction between ith machine and jth operator
- ϵ_{ijk} = random error

Assumptions:

- $\alpha_i \sim iidN(0, \sigma^2_\alpha)$
- $\beta_j \sim iidN(0, \sigma^2_\beta)$
- $(\alpha\beta) \sim iidN(0, \sigma^2_{\alpha\beta})$
- $\epsilon_{ijk} \sim iidN(0, \sigma^2)$
- $\alpha, \beta, (\alpha\beta), \epsilon$ all independent

NOTE:

field: 100198

field: ANOVA for two random factor model

field:

Source	DF	EMS
A	$a-1$	$\sigma^2 + n\sigma^2_{\alpha\beta} + nb\sigma^2_\alpha$
B	$b-1$	$\sigma^2 + n\sigma^2_{\alpha\beta} + na\sigma^2_\beta$
AB	$(a-1)(b-1)$	$\sigma^2 + n\sigma^2_{\alpha\beta}$
Error	$N-ab = ab(n-1)$	σ^2

- Balanced design $N = abn$
- To construct test for treatment X, find MS_X, and find EMS under H_0 for denominator in f test
- Interpretation: If we have significatn pvalue, there is evidence that response varies due to random effect A (if testing A)

NOTE:

field: 100199

field: Estimate variance components

field:

- Can either use MoM or REML (restricted maximum likelihood)
- For MoM, set MS sample quantities equal to their expectation (EMS) from ANOVA table
- Solve system of equations
- Note MoM estimates may not be in parameter estimates (ie variances negative, can just set to 0 if case), although this may indicate that model is inadequate
- If data are (approximately) balanced, and model is good, MoM and REML estimates should be close

NOTE:

field: 100200

field: Model and assumptions for 3 random factors design

field:

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

Where

- y_{ijkl}: strength of lth box from ith machine, jth operator and kth batch of glue
- μ = overall mean (fixed)
- α_i: effect of ith machine
- β_j: effect of jth operator
- γ_k: effect of kth batch of glue
- $(\alpha\beta)_{ij}$: machine $\times$ operator interaction
- $(\alpha\gamma)_{ik}$: machine $\times$ glue interaction
- $(\beta\gamma)_{jk}$: operator $\times$ glue interaction
- $(\alpha\beta\gamma)_{ijk}$: three way interaction
- ϵ_{ijkl}: random error

Assumptions

- Each random quantity X is independent from others and distributed iid $N(0, \sigma_X^2)$

NOTE:

field: 100201

field: Anova for 3 factor random effects

field:

- Needs approximate F test
- there is no MS with expectation of MS_X under H_0, so we must find a linear combination of the MS that has the right expectation: $\sum_s g_s MS_s$
- Since the denominator of F statistic is a linear combination of the MSs, the F test is approximate, so we have to approximate the degrees of freedom too
- Denominator df:

  $$v^* = \frac{(\sum_s g_s MS_s)^2}{\sum_s g_s^2 MS_s^2 / v_s}$$

  where v_s =df for MS_s (same as Satterthwaite approximation for Welch t-test)
- Generally dont estimate variance components (ie for a confidence interval), since these tests are asymptotic (unlike F -test)

NOTE:

field: 100202

field: Crossed factors: model and assumptions A is fixed B is random

field:

- Mixed effects model
- All combos of factors are tested
- Each machine is used by all operators
- Each operator produces boxes using both machines

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

- y_{ijk}: strength of kth box, made with ith machine and jth operator
- μ: overal mean (fixed)

- α_i: effect of ith machine (fixed)
- β_j: effect of jth operator
- $(\alpha\beta)_{ij}$: machine $\times$ operator interaction
- ϵ_{ijk}: random error

Assumptions

- $\sum_{i=1}^{2} \alpha_i = 0$
- $\beta_j iid N(0, \sigma_\beta^2)$
- $(\alpha\beta)_{ij} \sim iid N(0, \sigma_{\alpha\beta}^2)$
- $\epsilon_{ijk} \sim N(0, \sigma^2)$
- $\beta, (\alpha, \beta), \epsilon$ all independent

NOTE:

field: 100203

field: Nested factors : Model and Assumptions

field:

		B			
		1	2	3	4
A	1	$A_1B_{1(1)}$	$A_1B_{2(1)}$	$A_1B_{3(1)}$	$A_1B_{4(1)}$
	2	$A_2B_{1(2)}$	$A_2B_{2(2)}$	$A_2B_{3(2)}$	$A_2B_{4(2)}$

- A is fixed, B is random
- Each operator uses only one machine
- Neither machine is used by all operators
- Can't compare machine effect among operations(because we dont know how boxes would vary if the operator had used the other machine), so we cant model the interaction
- Model does not include $A \times B$

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$$

- y_{ijk}: strength of kth box, made with ith machine, and jth operator
- μ: overall mean
- α_i; effect of ith machine
- $\beta_{j(i)}$: effect of jth operator for the ith machine
- $\sum_{i=1}^{a} \alpha_i = 0$, $\beta_{j(i)} \sim iidN(0, \sigma_\beta^2), \epsilon_{ijk} \sim iidN(0, \sigma^2)$, both independent

NOTE:

field: 100204

field: Comparison between crossed factors and nested factors

field:

- Crossed factors: every level of A saw every level of B, and vice versa,
- Nested factors: if B is nested in A, levels of B only see one level of A

NOTE:

field: 100205

field: Reason for nesting

field:

- Feasibility, if machines are in different locations, we wouldnt want to transport operators around
- Subsampling: Multiple observations on the same experimental unit
  - Observation is nested in experimental unit
  - Experimental units are always nested in treatment (ie fish in temp fish tank), treatment tank, measurement fish
  - Usually nested effects are random, but not necessary

NOTE:

field: 100206

field: Multiple levels of nesting Model and assumption

field: A, factory (random), B, machine (random), C = operator (random)

- Operator is nested in machine version, nested in factory
- No crossed effects means no interaction term
- $y_{ijkl} = \mu + \alpha_i + \beta_{j(i)} + \gamma_{k(ij)} + \epsilon_{ijkl}$
- Assume each term iid Normal with associated variance, all independent.
- $Cov(y_{ijkl}, y_{ijkl'}) = \sigma^2_\alpha + \sigma^2_\beta + \sigma^2_\gamma$, covariance within same levels of each factor

NOTE:

field: 100207

field: Crossed and nested factors

field:

- All random
- Operator nested in machine (operators make boxes using only one machine
- Operator crossed with glue (operators make boxes using both batches of glue)
- So glue sees all operators and all machines
- $y_{ijkl} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{j(i)k} + \epsilon_{ijkl}$
- All normality and independence assumptions
- No three way interaction or machine times operator interaction

NOTE:

field: 100208

field: Estimating means and contrasts for a mixed model

field:

- A (machine) fixed, B (operator) random
- $V(\bar{y}_{1..}) = V\left(\frac{\sum_{j=1}^{b}\sum_{k=1}^{n} y_{1jk}}{bn}\right)\frac{1}{bn}(n\sigma_\beta^2 + n\sigma_{\alpha\beta}^2 + \sigma^2)$
- Calculated using independence assumptions, μ, α_1 fixed
- $\hat{C} = \sum_{i=1}^{g} w_i \bar{y}_{i.}$
- To compare machine 1 and 2, let $w_1 = 1$, $w_2 = -1$
- Compute point estimate $\hat{C} - \bar{y}_{1..} - \bar{y}_{2..}$

NOTE:

field: 100209

field: Complete Block Designs (RCB,RCBD)

field:

- Grouping observations into groups that are homogeneous
- Generalized pair - observations in a group not independent,
- Example: litter of animals, locations
- Experimental units stratified into blocks
- Within each block, randomly assign experimental units to treatments
- At least one replicate of each combination in each block

- In a balanced design, each block will have the same number of replicates for each treatment combination
- Draw out experimental design to identify blocked designs

NOTE:

field: 100210

field: Why use blocking

field:

- Account for non-independence
- Explain some of the variability in the response (blocking as a nuisence parameter)
- If experimental units can be grouped into homogeneous blocks, then blocks explain some of the variability
- variance reduction design

NOTE:

field: 100211

field: Model and assumptions for RCBD (with $n = 1$ observations per cell)

field:

- Resembles a factorial design
- $y_i = \mu + \alpha_i + \beta_j + \epsilon_{ij}$
- y_{ij}: response for the ith level of the treatment in the jth block
- μ: overall mean
- α_i: treatment effect

- β_j: block effect
- ϵ_{ij}; random error
- Assume: $\sum_i \alpha_i = \sum_j \beta_j = 0, \epsilon_{ij} \sim iidN(0, \sigma^2)$
- Note no interaction term - we dont want a blocking factor that interacts with treatments

NOTE:

field: 200212

field: ANOVA for RCBD ($n = 1$ replicate)

field: With a treatment with g gropus, and r blocks

Source	DF	SS
Treatment	$g-1$	
Block	$r-1$	
Error	$(g-1)(r-1)$	
Total	$gr-1=N-1$	

Note usually dont test block effect (but cant infer causation since not randomly assigned to blocks)

NOTE:

field: 200213

field: Relative efficiecy

field:

- Want to compare the amount of information captured from the data by two designs.
- Note a more complicated model (eg RCBD) would have a smaller SS_E but also a smaller df_{error}
- For a single observation from a normal distribution $I = \frac{1}{\sigma^2}$

- Information increases as variance decreases
- $RE = \frac{I_1}{I_2} = \frac{\sigma_2^2}{\sigma_1^2}$
- By convention: I_2 is the simpler design.
- Where σ_i^2 is the error variance in design i (which is assumed Normal)
- Since variances are not known, they must be estimated
- The variance of the design not performed will have to be calculated differently

NOTE:

field: 200214

field: Calculating and interpreting relative efficiency

field: EG for comparing CRD and RCBD

- $\widehat{RE} = \frac{\hat{\sigma}^2_{CRD}}{\hat{\sigma}^2_{RCBD}} \cdot \frac{(v_{RCBD}+1)(v_{CRD}+3)}{(v_{RCBD}+3)(v_{CRD}+1)}$
- Where v_{design} is the degrees of freedom for that design ($v_{CRD} = N - g$)
- $\hat{\sigma}_{CRD} = \frac{(r-1)MS_{block}+((g-1)+(r-1)(g-1)MS_E)}{(r-1)+(g-1)+(r-1)(g-1)}$ Weighted average of MS_{block} and MS_E
- IF RE = 2, then the RCBD is twice as efficient as CRD, so we should only need half as many replicates in the blocked design.

NOTE:

field: 100215

field: Latin squares design

field:

- If we have multiple blockign factors, this requires many eus.
- For a RCBD need at least one experiment unit in each cell for each treatment
- LS design is incomplete block design
- A Latin square design has g levels of the treatment, and 2 blocking factors, each with g levels. Each treatment level occurs exactlky once for each level of the blockign factor. (like sudoku)
- To randomize a LS experiment pick one LS at random from all possible LS designs of appropriate size. For $g = 3$, there are 12

NOTE:

field: 100216

field: Model and assumptions for Latin squares design

field:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk}$$

- y_{ijk} response for ith level of treatment, jth row(level of blocking factor 1), kth column (level of blocking factor 2)
- $\mu =$ overall mean
- α_i effect of level i of treatment
- β_j effect of level j of blocking factor (row effect)
- γ_k effect of level k of blocking factor 2 (column effect)
- Assumptions: Sum to zero constraints or normality/independence constraints if fixed/random. $\epsilon_{ijk} \sim N(0, \sigma^2)$

NOTE:

field: 100217

field: ANOVA for Latin Squares design

field:

Source	DF	MS	F
Treatment	(g-1)	MS(treat)	MS(treat)/MS(E)
Row	(g-1)	MS(row)	
Column	(g-1)	MS(col)	
Error	(g-1)(g-2)	MS(E)	
Total	$g^2 - 1 = N - 1$		

NOTE:

field: 100218

field: Relative efficiency of LS design compared to RCBD

field:

$$\widehat{RE}_{LS,RCBD} = \frac{\hat{\sigma}^2_{RCBD}}{\hat{\sigma}^2_{LS}} \frac{(v_{LS}+1)(v_{RCBD}+3)}{(v_{LS}+3)(v_{RCBD}+1)}$$

- Where $\hat{\sigma}_{RCBD} = \frac{(g-1)MS(row)+((g-1)+(g-1)(g-2)MS(E))}{2(g-1)+(g-1)(g-2)}$

NOTE:

field: 100219

field: Split-plot design description

field:

- Choose a split-plot design when some factors are more difficult or expensive to vary than others
- Example A - irrigation (large plot) and B - variety (small subplot of each large plot)
- Randomize assignment of levels of A

- In each level of A, randomize the subplots and assign levels of factor B
- Whole plots are experimental units wrt A, subplots are eus wrt B
- n - number of replicates for each level of A. If we want to replicate, we need an entire other set of levels of A.
- Assume whole plots are independent of each other
- Observations within a single whole plot are not assumed independent
- Whole plots nested in A, A and B are crossed

NOTE:

field: 100220

field: Model and assumptions for split plot design

field: example when fixed factors

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

- α_i effect of ith level of A
- $\eta_{k(i)}$ whole plot error , random effect for whole plot nested in A
- β_j effect of jth level of B
- $(\alpha\beta)_{ij}$ interaction of irrigation and variety
- $\epsilon_{k(i)j}$ random subplot error
- α, β, interaction assumptions based on fixed/random
- $\epsilon_{k(i)j} \sim iidN(0, \sigma^2)$ independent from $\eta_{k(i)} \sim N(0, \sigma_\eta^2)$
- Note that the first terms are a CRD model, which reflects we did a CRD at whole-plot level, if we did a more complicated design at whole-plot level, our model would reflect that.

NOTE:

field: 100221

field: Difference between split-plot design and complete factorial, blocking

field:

- Between complete factorial
  - Split plot has two stages of randomization
  - Complete factorial has one stage where combinations of factors are assigned to units
- Blocking
  - Randomization occurred at two stages, RDBD units are not randomly assigned to blocks
  - Dont care about inference for blocks, do care about inference for whole plot factor
- Nesting
  - Subplots are nested in whole plot, each level of A sees all level of B and each level of B sees all levels of A
  - Includes an interaction for A and B

NOTE:

field: 100222

field: ANOVA for split-plot design

field:

Source	DF	EMS
A	a-1	$\sigma^2 + b\sigma_\eta^2 + nb\frac{\sum_i \alpha_i^2}{a-1}$
Whole plot error	$a(n-1)$	$\sigma^2 + b\sigma_\eta^2$
B	$b-1$	$\sigma^2 + na\frac{\sum_j \beta_j^2}{b-1}$
AB	$(a-1)(b-1)$	$\sigma^2 + n\frac{\sum_i \sum_j (\alpha\beta)_{ij}^2}{(a-1)(b-1)}$
Subplot error	$a(n-1)(b-1)$	σ^2

NOTE:

field: 100223

field: Hypothesis tests for a split-plot design

field:

- For fixed effects,
- A effect: $H_0 : \alpha_i = 0$: $F = \frac{MS_A}{MS_{\text{whole plot}}}$
- B effect: $H_0 : \beta_j = 0$: $F = \frac{MS_B}{MS_{\text{sub plot}}}$
- AB effect $H_0 : (\alpha\beta)_{ij} = 0, F = \frac{MS_{AB}}{MS_{\text{subplot}}}$
- Under the null hypothesis, use the EMS for treatment factor to find the associated denominator EMS

NOTE:

field: 100224

field: Dependence between subplots of same whole plot - Cov and Corr

field:

$$\begin{aligned}
Cov(y_{ijk}, y_{ij'k}) &= Cov(\eta_{k(i)} + \epsilon_{k(i)j}, \eta_{k(i)} + \epsilon_{k(i)j'}) \\
&= \sigma_\eta^2 \\
Corr(y_{ijk}, y_{ij'k}) &= \frac{Cov(y_{ijk}, y_{ij'k})}{\sqrt{Var(y_{ijk})Var(y_{ij'k})}} \\
&= \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma^2}
\end{aligned}$$

NOTE:

field: 100225

field: Split plot design, estimate difference and CI (example)

field:

- Eg, CI for difference in 2 factors of B : $\beta_1 - \beta_2$: $\bar{y}_{\cdot 1 \cdot} - \bar{y}_{\cdot 2 \cdot}$
- Calculate variance by using model equation to find average, and use independence, and fixed variance rules.
- Replace MS_X as the estimate for σ_X^2
- Use t as reference distribution with DF for MS that gives us the variance estimate.

NOTE:

field: 100226

field: Repeated measures

field:

- Multiple observations on the SAME experimental unit
- Often repeated measurements in time.
- Questions to ask to use repeated measure
  - Ignoring repeated measurements, what is experimental design?
  - Are observations on the same EU independent?
  - What are research objective? (ie does it want an interaction between time and treatment)
- Use split-plot design, but time is not randomized

NOTE:

field: 100227

field: Repeated measures model and assumptions

field: Example model (fixed effects):

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

- μ: overall mean
- α_i: formula effect (fixed)
- $\epsilon_{k(i)}$: random baby effect
- β_j: time effect
- $(\alpha\beta)_{ij}$: formula times time interaction
- $(\epsilon\beta)_{k(i)j}$; baby $\times$ time interaction
- All assumptions for fixed/random factors (Here: $0 = \sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij}$)
- $\epsilon_{k(i)} \sim iidN(0, \sigma^2)$ independent from $(\epsilon\beta)_{k(i)j} \sim iidN(0, \sigma^2_{\alpha\beta})$
- Note assumes constant correlation over time, Correlation matrix R of one EU is 1 on diagonal and ρ everywhere else Correlation matrix overall is blocks of R on diagonal, 0 everywhere else.

NOTE:

field: 100228

field: ANCOVA

field:

- Same as regression parametrization for interaction terms - seperate intercepts equal slopes
- $y_{ij} = \mu + \alpha_i + \beta x_{ij} + \epsilon_{ij}$
- Assumes relationships between x_{ij} and y_{ij} is the same for all three groups (equal slopes) Or we could have had β_i to have different slopes
- Cant compare $\bar{y}_{1\cdot}, \bar{y}_{2\cdot}, \bar{y}_{3\cdot}$ since $\bar{x}_1 \neq \bar{x}_2 \neq \bar{x}_3$ (ie if we draw over mean y for mean x_i) these arent comparable.
- Compare mean ys by having common x value (often $\bar{x}$)
- Covariate adusted means: $y_{ij} = \tilde{mu} + \alpha_i + \beta(x_{ij} - \bar{x}_{\cdot\cdot}) + \epsilon_{ij}$
- $\tilde{\mu} - \beta\bar{x}_{\cdot\cdot} = \mu$