Exam 1 Notes

Chapter 1

- **Categorical**: groups/categories (e.g., hair color)
- **Quantitative**: numerical values
  - Discrete: countable set (# of siblings)
  - Continuous: value within some interval (call time on hold)
- Population = parameter | Sample = statistic
- Explanatory = x | Response = y

Chapter 2

- **Describing a Distribution**
  - 1. **Shape**
  - 2. **Center**
    - `> mean(setName)`
    - `> median(setName)`
    - `> sort(setName) - for mode`
  - 3. **Spread**
    - `> fivenum(setName)`
      - IQR = Q3 - Q1
    - `> quantile(setName)`
      - Rank/order: > (n * Percentile) + 0.5
    - `> sd(setName)`
    - `> var(setName) OR > sd(setName) ^ 2`
  - 4. **Outliers**: anything outside (Q1 - 1.5IQR) / (Q3 + 1.5IQR)
- **Graphs**
  - **Categorical**
    - FIRST create a **table**: `> tableName = table(setName)`
      - **Bar graph**: `> barplot(tableName)`
      - **Pie chart**: `> pie(tableName)`
  - **Quantitative**
    - `> plot(x,y)`
    - `> boxplot(setName horizontal = T)`
    - `> hist(setName)`
    - `> stem(setName)`

Chapter 3

- **Repeated values allowed**: n^r
- **Permutations**: ORDER is important
  - P^n_r = (n!) / (n-r)!
  - `> factorial(n)`
- **Combination**: unordered
  - `> choose(n, r)`
- **Relative Frequency**
  - P(E) = #(E elements) / (n observations)
- **Probability Rules**
  - **0 ≤ P(E) ≤ 1** for each event E
  - **P(Ω)** = 1 (sample space)
  - **P(∅)** = 0
  - **P(A)** = P(A∩B) + P(A∩~B)
  - **P(A)** = P(A|B1) * P(B1) + P(A|B2) * P(B2) + …
  - **Complement**: P(A∩~B) = P(A) - P(A∩B)
    - P(~A) = 1 - P(A)
    - P(~(A∪B)) = 1 - (A∪B)
  - **Addition**: P(A∪B) = P(A) + P(B) - P(A∩B)
  - **Multiplication**: P(A∩B) = P(A) * P(B|A) = P(B) * P(A|B)
  - **Conditional**: P(A|B) = P(A∩B) / P(B) = P(A) * P(B|A) / P(B)
- **Disjoint (Mutually Exclusive) vs Independent**
    - **Disjoint**: P(A∩B) = 0
    - **Independent**: P(A) = P(A|B)
      OR P(B) = P(B|A)
      OR P(A∩B) = P(A) * P(B)

Chapter 4 - Distribution

- **E[X]** = `> sum(x*y)`
- **E[X^2]** = `> sum((x^2)*y)`
- **Variance = var[X]** = sd^2 = E[X^2] - $E[X]^2$
- **sd_x** = sqrt(var[X])
- Probability values of P(X)/f(x) should add up to 1
- **Binomial**: (n - trials, p - prob of success)
  - X ~ binomial(n,p)
  - **P(X=x) =** `> dbinom(x,n,p)`
  - **P(X<=x) =** `> pbinom(x,n,p)`
  - **P(X>x) =** `> 1 - pbinom(x,n,p)`
  - Note (will be on given formula sheet)
    - **μ/Mean = E[X] = n*p**
    - **σ^2/Variance = Var[X] = np(1-p)**
- **Possion**: (μ - mean/avg)
  - X ~ poisson(μ)
  - **P(X=x) =** `> dpois(x,μ)`
  - **P(X<=x) =** `> ppois(x,μ)`
  - **P(X>x) =** `> 1 - ppois(x,μ)`
- **Hypergeometric**: (m - # success, n - # fails, k - *sample* size)
  - X ~ hyper(m,n,k)
  - **P(X=x) =** `> dhyper(x,m,n,k)`
  - **P(X<=x) =** `> phyper(x,m,n,k)`
  - Ex: X ~ hyper(m=20,n=15,k=5)
    P(X>=2) = 1 - P(X<=1) = 1 - phyper(1,20,15,5)

Chapter 9 - LSLR

- **Set X & Y**:
  - `> x = c(2,8,8,13,16,19)`
  - `> y = c(22,29,28,40,33,41)`
- **Scatterplot**: > plot(x,y)
- **Correlation Coefficient - r**: `> cor(x,y)`
- **Coefficient of Determination - r^2**: `> cor(x,y) ^ 2`
- **LSLR**: `> xy.lm = lm(y~x)`
  `> summary(xy.lm)`
  - **ŷ** = intercept + slope*x
  - **slope** = cor(x,y) * (sd(y)/sd(x))
  - **intercept** = mean(y) - slope*mean(x)
- **Residual**: `> summary(xy.lm)`
  - look at residual section of summary(lm)
  - **residual** = observed y - predicted y
- **Is it a good model?**
  - r^2 > 0.8, GOOD
  - r^2 < 0.5, NOT GOOD

Exam 2 Notes

Chapter 5 (Lec 9)

- **Types of Random Variables (Quantitative)**
  - Discrete: countable set (finite or infinite sequence)
  - Continuous: value within some interval
- **Probability Distribution**
  - **Discrete**: **probability mass function (pmf)**
    - Provide probability for EACH VALUE
    - **pmf**, $f(x) = P(X = x)$
  - **Continuous**: **probability density function (pdf)**
    - Graph of an equation within an INTERVAL
    - **pdf**, $f(x) \neq P(X = a) = {}_a\int^a f(x)dx = 0$ for all x
      $f(x) = P(a \leq X \leq b) = {}_a\int^b f(x)dx$
    - Note: ${}_{-\infty}\int^{\infty} f(x)dx = 1$
    - **cdf**, $F(x) = P(X \leq x)$ → just plug in x into F(x)
- **Uniform Distribution**
  - **pdf** of X is: $f(x) = 1 / (B-A), \quad A \leq x \leq B$
    $0, \quad$ otherwise
  - **cdf** of X is: $F(x) = 0, \quad x < A$
    $(x - A) / (B-A), \quad A \leq x \leq B$
    $1, \quad x > B$
- **Using cdf F(x) for Probabilities**
  - $P(X > a) = 1 - F(a)$
  - $P(a \leq X \leq b) = F(b) - F(a)$
- **cdf to pdf**: $F'(x) = f(x)$ (pdf = derivative of cdf)

Chapter 5 (Lec 10)

- **Expected Values** (Continuous Random Variables)
  - $E(X) = {}_{-\infty}\int^{\infty} xf(x)dx$
  - $E(h(X)) = {}_{-\infty}\int^{\infty} h(x)f(x)dx$
- **Exponential Distribution**
  - **pdf**, $f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$
    $0, \quad x < 0$
  - **cdf**, $F(x) = 1 - e^{-\lambda x}, \quad x \geq 0$
    $0, \quad x < 0$
  - Mean / $\mu_x = E(X) = 1/\lambda$
  - St dev = $1/\lambda$
  - $Var(X) = (1/\lambda)^2 = 1/\lambda^2$
  - **X~exp(λ (= 1/μ) = ?)**
    - `P(X ≤ x): > pexp(x, λ)`
    - `Percentile: > qexp(x, λ)`
- **Gamma Function**
  - $\Gamma(\alpha) = {}_0\int^{\infty} x^{\alpha-1}e^{-x}dx$
  - Properties
    - For any $\alpha > 1$, $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$
    - For any positive int n, $\Gamma(n) = (n-1)!$
    - $\Gamma(1/2) = \sqrt{\pi}$
- **Gamma Distribution**
  - **pdf**, $f(x; \alpha,\beta) = (x^{\alpha-1}e^{-x/\beta}) / \beta^{\alpha}\Gamma(\alpha) \quad x \geq 0$
    $0 \quad$ otherwise / $x < 0$
  - **X~gamma(α,β)**
    - `P(X ≤ x): > pgamma(x,α,1/β)`
  - Note: ***if*** $\alpha = 1$ $\quad f(x; \alpha,\beta) = (e^{-x/\beta}) / \beta$, **X~exp(λ = 1/β)**
  - $E(X) = \mu = \alpha\beta$
  - $Var(X) = \sigma^2 = \alpha\beta^2$
- **Normal Distribution**
  - **pdf**, $f(x) = e^{-(x-\mu)^2 / 2\sigma^2} / \text{sqrt}(2\pi)\sigma$
  - **X~N(μ,σ), E[X] = μ, sd(x) = σ**
    - `P(X ≤ x): > pnorm(x,μ,σ)`
  - **Empirical Rule (68-95-99.7)**
    - $P(\mu-1\sigma < X < \mu+1\sigma) = 0.68$
    - $P(\mu-2\sigma < X < \mu+2\sigma) = 0.95$
    - $P(\mu-3\sigma < X < \mu+3\sigma) = 0.997$

Chapter 5 (Lec 11)

- **Standard N.D. Z-score:** # of standard deviations from mean
  - The larger the |z| value, the more "unusual"
  - $Z = (X - \mu) / \sigma$
  - $E[Z] = 0$
  - $\sigma(Z) = 1$
  - **Z~N(μ=0,σ=1)**
    - `P(Z ≤ x): > pnorm(x,0,1)`
    - or refer to z-score table
- **Inverse Normal**: finding obs value when given proportion
  - **Z~N(μ=0,σ=1)**
    - `P(Z ≤ x): > qnorm(proportion,0,1)`
  - **X~N(μ,σ)**
    - `P(X ≤ x): > qnorm(proportion,mean,sd)`
- **Binomial With Normal Distribution**
  - $\mu = np$
  - $\sigma = \text{sqrt}(np(1-p))$
  - X~Binom(n,p): n trials, p probability of success
    - `P(X ≤ x): > pbinom(x,n,p)`
  - **X~N(μ=np,σ=sqrt(np(1-p)))**
    - `P(Z ≤ x): > pnorm(x+0.5,μ,σ)`
- **Recall**
  - $\mu_{X+Y} = E[X+Y] = E[X] + E[Y] = \mu_X + \mu_Y$
  - $\mu_{X-Y} = E[X-Y] = E[X] - E[Y] = \mu_X - \mu_Y$
  - INDEPENDENT X & Y
    - $\sigma^2_{X+/-Y} = Var[X+/-Y] = Var[X] +/- Var[Y] = \sigma^2_X +/- \sigma^2_Y$
  - DEPENDENT X & Y
    - $\sigma^2_{X+Y} = Var[X+Y] = \sigma^2_X + \sigma^2_Y + 2cov(X,Y)$
    - $\sigma^2_{X-Y} = Var[X-Y] = \sigma^2_X - \sigma^2_Y - 2cov(X,Y)$

Chapter 6 (Lecture 12)

- **Sampling Distribution (for sample mean x̄)**
  - Characteristics:
    - Shape, center, spread
  - $\mu_{\bar{x}} = \mu = E[\bar{x}]$
  - $\sigma_{\bar{x}} = \sigma / \text{sqrt}(n)$
  - $Var[\bar{x}] = \sigma^2 / n$
  - $\mathbf{z = (\bar{x}-\mu) / (\sigma/\text{sqrt}(n))}$
- **Shape: if population ~ N(μ,σ), THEN x̄ ~ N(μ,σ/√n)**
  - **Central limit theorem**: if we don't know about population, as long as (n > 30) then we can assume $\bar{x} \sim N(\mu,\sigma/\sqrt{n})$ [use CLM to assume shape is normal]
  - **x̄ ~ N(μ,σ/√n)**
    - `P(x̄ ≤ x): > pnorm(x,μ,σ√n)`
- **Sample Proportions (p̂)**
  - $\hat{p} = X / n$ where x # of success, n # of obs (sample size)
  - $\mu_{\hat{p}} = E(\hat{p}) = p$
  - $\sigma_{\hat{p}} = \text{sqrt}(p(1-p) / n)$
  - $\sigma^2_{\hat{p}} = Var(\hat{p}) = p(1-p) / n$
  - **10% Condition**: rand and ind when samp size ≤ 10% pop
  - **Success/Failure Condition**: normal distribution IF successes (np) ≥ 10 & fails (n(1-p)) ≥ 10
  - $\hat{p} \sim N(\mu, \sigma=\text{sqrt}(p(1-p) / n))$
    - `P(x̄ ≤ x): > pnorm(x,μ,σ)`

Chapter 7 (Lec 13-14) Confidence Interval

- **Statistical Inference**
  - Estimation & hypothesis testing (NOT mutually exclusive)
    - $E(\bar{x}) = \mu$ → unbiased estimator
    - $E(\hat{p}) = p$ → unbiased estimator
    - $E(s^2) \neq \sigma^2$ → biased estimator
- **Standard Error**: SE(SE(θ^) = sqrt(Var(θ^))
  - SE($\bar{x}$) = σ/sqrt(n)
  - SE($\hat{p}$) = sqrt(p(1-p) / n)
- **Confidence Interval**
  1. Get level of confidence
  2. Compute margin of error
  3. Interpret: We are "_"% confident that the "population parameter" is between "lower limit" and "upper limit"
  - NOTE: higher CI → wider interval/larger M.E.
- **Z-Distribution**: ***σ is KNOWN – CI for μ***

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

  - $\bar{X}$: sample mean
  - $z_{\alpha/2}$: critical value
  - 1- α: confidence level
  - σ: population standard deviation
  - n: sample size
  - $z_{\alpha/2}(\sigma/\sqrt{n})$: margin of error
  - `CI: > xbar + c(-1,1) * qnorm(...) * σ/√n`
    - **Critical value**: $z_{\alpha/2}/z^*$ = `> qnorm((1+C)/2)`
    - **Margin of Error**: m/me = critical value * SE
      - m = $z_{\alpha/2}$*(σ/√n) OR m = width/2
- **T-Distribution**: ***σ is UNKNOWN (only sd from sample)***

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

  - Depends on degrees of freedom (df = n-1)
  - `CI: > xbar + c(-1,1) * qt(...) * s/√n`
    - **Critical value**: t = `> qt((1+C)/2, df)`
    - t = ($\bar{x}$-μ) / (s/sqrt(n))
- **Sample Size (based on CI and Mean)**

$$n > \left(\frac{z_{\alpha/2}\sigma}{m}\right)^2$$

- **Proportions: CI for proportions/percentages**
  - Conditions:
    - Population must be ≥ 10 times size of sample
    - #successes ($n\hat{p}$) ≥ 10 & #fails ($n(1-\hat{p})$) ≥ 10

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

  - `CI: > p+c(-1,1)*qnorm(...)*sqrt(p*(1-p)/n)`
- **Sample Size (based on CI and Proportion)**

$$n > p^*(1-p^*)\left(\frac{z_{\alpha/2}}{m}\right)^2$$

- **Distribution for Variance/SD: chi-square distribution**
- **$X^2 = (n-1)s^2 / \sigma^2$**
  - `P(X² ≤ x): > pchisq(x, df)`
  - `P(X² > c)= x: > qchisq(1-x, df)`
- **Confidence Interval - Chi-Square (Var & SD)**
  - **CI for s^2 = variance**
  - `lcl: ((n-1)*s²)/qchisq(1-(α/2), n-1)`
  - `ucl: ((n-1)*s²)/qchisq(α/2, n-1)`
  - **CI for Standard Deviation**
  - `lcl: sqrt[((n-1)*s²)/qchisq(1-(α/2), n-1)]`
  - `ucl: sqrt[((n-1)*s²)/qchisq(α/2, n-1)]`

Chapter 8 (Lec 15) Hypothesis Test

- **Hypothesis/Significance Test**
  1. **Check Assumptions**
     - An SRS of size n from the population
     - Z-test (know σ) OR T-test (unknown σ)
     - Either a Normal pop. or a large sample (n ≥ 30)
  2. **State Null Hypothesis (H_0) & Alternative Hypothesis (H_a)**
     - **H_0**: μ **=** "value" assumed to be true
     - **H_a**: μ **≠** "value" assumed to be true
       - **left-tailed** test - **H_a**: μ **<** μ_0
       - **right-tailed** test - **H_a**: μ **>** μ_0
       - **two-tailed** test - **H_a**: μ **≠** μ_0 ("different")
  3. **Rejection region** (graph & label critical value)
     - SKIP IF ALPHA IS NOT GIVEN
     - LTT: μ < μ_0, reject region is in the left tail
       - CV: `> qnorm(α) OR qt(α, n-1)`
       - Reject H_0 if z ≤ CV
     - RTT: μ > μ_0, reject region is in the right tail
       - CV: `> qnorm(1-α) OR qt(1-α, n-1)`
       - Reject H_0 if z ≥ CV
     - TTT: μ ≠ μ_0, reject region is in both tails
       - CV: `> qnorm(α/2)/qt(α/2, n-1)` `qnorm(1-(α/2))/qt(1-(α/2), n-1)`
       - Reject H_0 if z ≥ CV || z ≤ -CV
  4. **Calculate Test Statistic (z-stat or t-stat)**
     - Used to measure the difference between the data and what is expected on the null hypothesis
       (σ is known) (σ is NOT known)

$$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \qquad t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

  5. **Find P-value (probability)**
     - Based on significance level (α)
       - H_a: μ < μ_0, then P-value = P(Z < test statistic)
       - H_a: μ > μ_0, then P-value = P(Z > test statistic)
       - H_a: μ ≠ μ_0, then P-value = 2P(Z < test statistic)
     - Reject H_0 if P-value ≤ α (can say that the data is statistically significant at level α)
     - Fail to reject H_0 if P-value > α
  6. **Conclusion: 2 Possible Decisions of Test**
     - Reject H_0 in favor of H_a (RH0)
       - There is some/strong/very/extremely …
     - Fail to reject null hypothesis (FTRH0)
       - There is no evidence that …
     - *(NEVER accept null hypothesis)*
     - Conclude in context of problem w confidence of _%
- **Decision Errors:**

Our Decision	Correct Condition	
	H_0 is true	H_0 is false
Reject H_0	Type I Error	Correct
Fail to reject H_0	Correct	Type II Error

  - P(Type I Error) = α
  - P(Type II Error) = β
  - Power = 1 - β
- **Not Given α**
  - If the P-value for testing H_0 is less than … (reject H_0)
    - P < 0.1: some evidence that H_0 is false
    - P < 0.05: strong evidence that H_0 is false
    - P < 0.01: very strong evidence that H_0 is false
    - P < 0.001: extremely strong evidence that H_0 is false
  - If the P-value is greater than 0.1, we do not have any evidence that H0 is false (fail to reject H_0)

Chapter 8 (Lec 15) Hypothesis for Proportions

- **Hypothesis**
  - **H_0**: p **=** p_0
  - **H_a**: p **≠** p_0
    - **left-tailed** test - **H_a**: p **<** p_0
    - **right-tailed** test - **H_a**: p **>** p_0
    - **two-tailed** test - **H_a**: p **≠** p_0 ("different")
- **Conditions**
  - Sample must be an SRS from the population of interest
  - Population must be at least 10 times the size of the sample
  - Number of successes and the number of failures must each be at least 10 (both $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$).
- Note: $\hat{p}$ = # of successes/# of observations = x/n
- Use **z-test statistic**: z = ($\hat{p}$ - p_0) / sqrt(p(1-p) / n)
- `> prop.test(x=575, n=1000, p=0.5, alternative="greater", correct=F)`
  - x: # of successes, n: sample size, p: null hypothesis, alternative = c("two.sided", "less", "greater"),

Significance Test Summary

Parameter	μ given σ	μ **not** given σ	p proportions
1. Null hypothesis	$H_0 : \mu = \mu_0$		$H_0 : p = p_0$
2. Alternative	Choose either $<$, $>$, or $\neq$ in place of $=$ in H_0.		
3. Rejection Region Depending on H_a.	$z_{\alpha/2}$	$t_{\alpha/2}$ with df = n - 1	$z_{\alpha/2}$
4. Test statistic	$z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$	$t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$	$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
5. P-value	pnorm(z)	pt(t,n-1)	pnorm(z)
This is the area under the density curve shaded according to H_a.			
6. Decision	**Reject** H_0 if P-value $\leq \alpha$ **Fail** to reject H_0 if P-value $> \alpha$		

Chapter 10 (Lec 16) Inferences on 2 Groups

- **Matched Pairs t-Test**
  - Data samples are DEPENDENT of each other
  - Hypothesis:
    - **H_0**: μ_d **=** 0 & **H_a**: μ_d **≠** 0, μ_d **>** 0, μ_d **<** 0
    - μ_d is the mean of differences
  - Confidence Interval: $\bar{x}_d$ ± t * (s_d / sqrt(n))
    - $\bar{x}_d$ = add differences / n
    - s_d = sd(differences)
    - t = qt((1+C)/2, n-1)
  - Rcode:
    - `> setA=c(1, 2, ...)`
    - `> setB=c(1, 2, ...)`
    - `> t.test(setA,setB,alternative="?", conf.level = ?, paired = TRUE)`
- **2 Two-Population Inference**
  - Data samples are INDEPENDENT of each other
  - Interval of Estimation:
    - Point Estimate: $\bar{x}_1 - \bar{x}_2$
    - Confidence level: 1 − α = C
    - Critical value: t∗ = qt((1+C)/2,df).
    - Margin of Error: E = t*sqrt(s_1^2/n_1 + s_2^2/n_2)
    - Confidence Interval: point estimate ± margin of error
    - CI = $\bar{x}_1 - \bar{x}_2$ + c(-1,1)*qt((1+C)/2, df)*sqrt(s_1^2/n_1 + s_2^2/n_2)
  - Conclusion: we are ?% confident that the difference in mean (subject)? of (setA) vs (setB) is between (LB) and (UB).
- **Two Sample t-Test (Comparing Two Means)**
  - Data samples are INDEPENDENT of each other
  - Hypothesis:
    - **H_0**: μ_1 = μ_2 & **H_a**: μ_1 **≠** μ_2, μ_1 **>** μ_2, μ_1 **<** μ_2
  - t = ($\bar{x}_1 - \bar{x}_2$) / sqrt(s_1^2/n_1 + s_2^2/n_2)
  - Rcode:
    - `> setA=c(1, 2, ...)`
    - `> setB=c(1, 2, ...)`
    - `> t.test(setA,setB,mu=0,alternative="?")`
  - Conclusion: There is some/strong/very/extremely evidence that mean (subject) are significantly different (setA) vs (setB)
- **Comparing Two Proportions**
  - X~Bin(n,p): $\hat{p}$ = (x/n) | E($\hat{p}$) = p | SD($\hat{p}$) = sqrt(p(1-p)/n)
  - Hypothesis:
    - **H_0**: p_1 **=** p_2 & **H_a**: p_1 **≠** p_2, μ_1 **>** p_2, p_1 **<** p_2
  - Interval of Estimation
    - Point Estimate: $\hat{p}_1 - \hat{p}_2 = \bar{x}_1/n_1 - \bar{x}_2/n_2$
    - Confidence level: 1 − α = C
    - Critical value: z∗ = qnorm((1+C)/2)

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

    - Confidence Interval: point estimate ± margin of error
    - CI = $\hat{p}_1 - \hat{p}_2$ + c(-1,1)*qnorm((1+C)/2, df) * sqrt($\hat{p}_1(1-\hat{p}_1)/n_1$ + $\hat{p}_2(1-\hat{p}_2)/n_2$)
  - Rcode:
    - `> prop.test(x=c(x1,x2), n=c(n1,n2), conf.level = C, correct = FALSE)`

Chapter 9 (Lec 17) LSRL

- **$Y = \beta_0 + \beta_1 x + \varepsilon$**
  - Y: dependent variable (response)
  - x: independent variable (explanatory)
  - β_0: population intercept
  - β_1: population slope
  - ε: error term
- **residual** = observed y - predicted y
- **LSLR**: `> xy.lm = lm(y~x)`
  `> summary(xy.lm)`
- **T Test Significance of β_1**
  - Hypotheses: **H_0**: $\beta_1 = 0$ // **H_a**: $\beta_1 \neq 0$
  - Test statistic: t = (β_1 - $\beta_{hypothesis}$) / sd
  - P-value: t distribution with n-2 degrees of freedom
    - Two-tailed: p-val = `> 2 * pt(-t, df)`
  - Decision: Reject H_0 if p-value ≤ α
  - Conclusion: If H_0 is rejected we conclude that explanatory variable x can be used to predict the response variable y
- **Confidence Interval for β_1**

$$b_1 \pm t_{\alpha/2,n-2} \times SE_{b_1}$$

  - t* (critical value): `> qt((1+C)/2, df)`
  - CI: `> confint(xy.lm, level=0.95)`

Chapter 11 (Lec 18) More Than 2 Means

- **More Than Two Means Test**
  - Question: is there a "*statistically significant difference*" in the mean (subject) among the n (groups)?
  - Null hypotheses: mean (subject) is same among n means
    - $H_0 : \mu_{group1} = \mu_{group2} = \mu_{groupn} = \ldots$
  - Alternative hypothesis: at least one of the mean (subject) among the *n* (groups) is different
  - Conclusion: rejecting H_0 is evidence that the mean of at least one group is different from the other means
- **Formulas**
  - Note:
    - $\bar{X}_i$ = group mean
    - $\bar{X}_{..}$ = grand mean
    - M = total # of groups
    - N = total # of observations
  - SSTr: treatment sum of squares (between groups)

$$SS(betw) = \sum_{i=1}^{M} n_i(\bar{X}_{i.} - \bar{X}_{..})^2$$

  - SSE: error sum of squares (residual)

$$SSE = SS(resid) = \sum_{i=i}^{M}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2 = \sum_{i=1}^{M}(n-1)S_i^2.$$

  - SST: total sum of squares

$$SS(tot) = \sum_{i=i}^{M}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{..})^2 = SS(betw) + SS(resid)$$

- **F Test**
  - Mean square for treatments is MSTr = SSTr / M-1
  - Mean square for error is MSE = SSE / N-M
  - Test statistic is F = MSTr / MSE
    - F distribution with parameters "numerator df" = M - 1 and "denominator df" = N - M
- **ANOVA Table (ANalysis Of VAriance)**
  -

Source of Variation	degrees of freedom	Sum of Squares	Mean Square	F
Treatments	M - 1	SSTr	MSTr	$\frac{MSTr}{MSE}$
Error	N - M	SSE	MSE	
Total	N - 1	SST		

    - p-value = `> 1 - pf(f, M-1, N-M)`
  - Generate ANOVA: `> anova(xy.lm)`

Chapter 12 (Lec 19) Chi-Square

- **Goodness of Fit Tests**
  - Tests how well sample proportions of categories "match-up" with the known population proportions
  - Hypotheses:
    - **H_0**: proportions are the same as what is claimed
    - **H_a**: at least one proportion is different than claimed
  - Test Statistic: chi-square

Observed Counts (O)	Expected Counts (E)	$\frac{(O-E)^2}{E}$

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

    - Expected count = POP TOTAL count * proportion
    - df = n-1
    - P-val (X^2 ≥ test stat): `> 1 - pchisq(x,df)`
    - Chisq: `> chisq.test(c(list of obs vals), p=c(list of props))`
  - Conclusion: fail to reject H_0, there is no evidence that the (subject) is difference from what (name) claims
- **X^2 Test of Independence (Significance Test)**
  - Hypotheses:
    - Null hypothesis: There is no association (independence) between row & column variables
    - Alternative hypothesis: There is an association (dependence) bt row variable and column variable
    - **H_0** : Airline & on-time performance are independent.
      H_A : On-time performance depends on airline.
  - Test Statistic: chi-square

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

    - Expected count = (row total * col total) / TOTAL n
    - df = (r-1)(c-1)
    - P-val (X^2 ≥ test stat): `> 1 - pchisq(x,df)`
    - Chisq:
```
> matrixName = matrix(c(...), nrow=?, ncol=?)
> matrixName        # use to see matrix
> chisq.test(matrix, correct = F)
```
  - Decision: if p-val less than α level of significance, we reject H_0 (dependence), otherwise fail to reject H_0(no association)