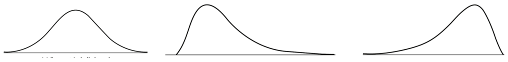
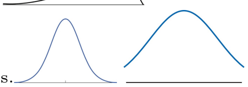
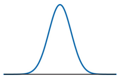
- **C0 Why Study Statistics? & C1 An Introduction to Statistics and Statistical Inference**
  - **Statistics** is science of collecting and interpreting data. Components: collection, organization, analysis, interpretation.
  - Branches of statistics: Collection of data, descriptive statistics (graphical and numerical methods used to describe, organize, summarize data), inferential statistics(techniques and methods used to analyze a small, specific set of data in order to draw a conclusion about a large, more general collection of data).
  - Inferential statistics. Claim (status quo), experiment (check claim), likelihood (consistent with claim?), conclusion (reasonable vs. rare).
  - **Population** is entire collection of individuals or objects to be considered or to be studied. **Sample** is a subset of population, small selection of individuals or objects taken from entire collection. **Variable** is characteristic of an individual or object in a population of interest (quantitative and qualitative).
  - **Solution Trial.** 1) Find keywords. 2) Correctly translate words into statistics. 3) Determine applicable concepts. 4) Develop vision or strategy for solution. 5) Solve problem.
- **C 2 Tables and Graphs for Summarizing Data**
  - **Variables**. Number (univariate, bivariate, multivariate) and type (numerical, categorical)
  - Ask: Who? (What/how many cases?) What? (How many variables, exact definition for each variable, unit of measurement for each variable?) Why (Purpose, questions being asked, suitable variables?)
  - Look for in graphs: shape, center, variability.
  - **Frequency distribution.** Label or class is category of data. Frequency is count. relative frequency $= \frac{\text{frequency}}{\text{total count}}$.
  - **Histograms** show distribution of quantitative variable by using bars. Procedure for DISCRETE: 1) Calculate frequency distribution and/or relative frequency of each value. 2) Mark possible values on x-axis. 3) Above each value, draw a rectangle whose height is the frequency (or relative frequency) of that value. Procedure for CONTINUOUS: 1) Divide x-axis into number of equal class intervals or classes such that each observation falls into exactly one interval. # classes $\approx \sqrt{\#\text{ observations}}$. 2) Calculate frequency or relative frequency for each interval. Above each value, draw a rectangle whose height is frequency or relative frequency of that value.
  - Examining distributions. In any graph of data, look for overall patterns and for striking deviations from that pattern. Describe overall pattern by shape, center, spread. Outlier is individual that falls outside overall pattern.
  - Symmetric, positively (right) skewed, negatively (left) skewed.

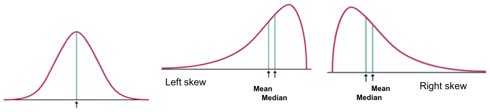


  - Normal distribution, heavy tails, light tails.

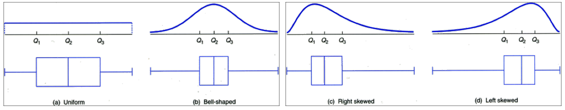




- **C3 Numerical Summary Measures**
  - Measures of central tendency indicate where majority of data is centered.
  - Lowercase Latin letters indicate variables. $x_1, x_2, \ldots, x_n$ refers to set of fixed observations of a variable. n is number of observations, sample size.
  - **Sample mean.** $\overline{x} = \frac{1}{n}\sum x_i$. μ is population mean. Sample: Latin letters, population: Greek letters.
  - **Sample median** $\tilde{x}$. Procedure: 1) Sort observations smallets to largest. n odd $\Longrightarrow$ $\tilde{x}$ is center. n even $\Longrightarrow$ $\tilde{x}$ is average of two center observations.
  -



  - **Mode** M is value with greatest frequency.
  - range = maximum − minimum.
  - **Variance**. $s_x^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2$. Standard deviation $s_x = \sqrt{\frac{1}{n-1}\sum(x_i - \overline{x})^2}$. σ^2 is population variance. S.D. is used to determine spread for comparisons. $s^2 = 0 \Longrightarrow$ all observations are the same, $s > 0$. $n = 1$ means no S.D. s has same units of measurement as does original observations.
  - **Quartiles**. Procedure: 1) Sort values from lowest to highest, locate median. Q_1 is median of lower half. $d_1 = n/4$. d_1 integer $\Longrightarrow$ Q_1 is mean of observations at d_1 and $d_1 + 1$. Not d_1 integer $\Longrightarrow$ Q_1 is observation at $\lceil d_1 \rceil$. 3) Q_3 is median of upper half. Compute $d_3 = 3n/4$ and repeat.
  - $IQR = Q_3 - Q_1$
  - **Outliers.** $\text{IF}_\text{L} = Q_1 - 1.5(\text{IQR})$. $\text{IF}_\text{H} = Q_3 + 1.5(\text{IQR})$. $\text{OF}_\text{L} = Q_1 - 3(\text{IQR})$. $\text{OF}_\text{H} = Q_3 + 3(\text{IQR})$.
  - **Five-number summary.** Minimum, Q_1, median, Q_3, Maximum.
  - **Boxplots.** Procedure: 1) Find $Q_{1,3}$, median, IQR. 2) Calculate fences. Draw central box from Q_1 to Q_3. Draw line for median. Extend whiskers from box to minimum and maximum values that are NOT outliers. 4) Closed circles for mild outliers, open circles for extreme outliers.
  -



  - If data is mostly symmetrical, use mean and standard deviation. When data is skewed, use median and IQR.
  - **Empirical rule.** 68% of data is within one S.D. of mean. 95% of data is within two S.D.s of mean. 99.7% of data is within three S.D.s of mean.
  - **z-score**. $z_i = \frac{x_i - \overline{x}}{s}$. Measure of relative standing. Given a set of n observations, sum of z-scores is 0.
- **§4.1 Experiments, Sample Spaces, and Events**
  - **Experiment**: activity in which there are at least 2 possible **outcomes** (results of experiment), result cannot be predicted w/ absolute certainty; **trial**: experiment done once.
  - **Sample space**: listing of all possible outcomes in experiment, denoted by S or Ω.
  - **Event**: any collection of outcomes from experiment; **simple event**: exactly one outcome; event has **occurred** if resulting outcome is contained in event
  - For events A, B: A' is A complement (NOT); $A \cup B$ is A OR B OR both; $A \cap B$ is BOTH A AND B.
  - A, B are **disjoint** or **mutually exclusive** if nothing in common; $A \cap B = \{\}$.
- **§4.2 An Introduction to Probability**
  - **Probability** of outcome of chance process is proportion of # times outcome would occur in series of repetition; $P(A) \approx \lim_{\#\text{ trials}\to\infty} \frac{\#\text{ times } A \text{ occurs}}{\#\text{trials}}$.
  - Bayesian probability belongs to the category of evidential probabilities; to evaluate the probability of a hypothesis, the Bayesian probabilist specifies some prior probability, which is then updated in the light of new, relevantdata(evidence).
  - For any event A, $P(A)$ is sum of probabilities of all outcomes in A. $P(S) = 1$. $P(\{\}) = 0$.
  - Types of probability: Subjective (no math), empirical ($P(A) = \frac{\#\text{ times } A \text{ occurs}}{\text{total }\#\text{ times}}$), theoretical/equally likely ($P(A) = \frac{\#\text{ outcomes in } A}{\#\text{ outcomes in } S}$).
  - For any A, $P(A') = 1 - P(A)$.
  - **Addition Rule.** For any A, B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. For any DISJOINT A, B, $P(A \cup B) = P(A) + P(B)$.
- **§4.4 Conditional Probability & §4.5 Independence**
  - $P(A \mid B) = \frac{P(A\cap B)}{P(B)}$.
  - Two events A, B are independent $\iff$ $P(A \mid B) = P(A)$. Two events are independent if knowing that one occurs does not change the probability that the other occurs. If A, B are two independent events, then so are all combinations of these two events and their complements, and $P(A \cap B) = P(A)P(B)$.
  - **General Multiplication Rule.** $P(A \cap B) = P(A)P(B \mid A)$. $P(A \cap B \cap C) = P(A)P(B \mid A)P(C \mid A \cap B)$.
  - **Bayes' Rule (for two variables).** $P(A \mid B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A')P(A')}$.
  - **Bayes' Rule.** Suppose a sample space is decomposed into k disjoint events $A_1, A_2, \ldots, A_k$, such that $P(A_i) > 0$ and $\sum_{i=1}^k P(A_i) = 1$. Then $P(A_j \mid B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^k P(B|A_i)P(A_i)}$.
- **§5.1 Random Variables & §5.2 Probability Distributions for Discrete Random Variables**
  - **Random variable**: a function that assigns a unique numerical value to each outcome in a sample space.
  - **Probability distribution of a random variable** gives all of its possible values and the probabilities of each of them.
  - **Probability mass function (pmf)** is probability that a D.R.V. is equal to some specific value. $p(x) = P(X = x)$

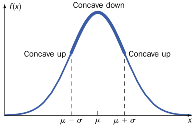
Outcome	x_1	x_2	...
Probability	p_1	p_2	...

  - Properties of a valid probability distribution. 1) $0 \le p_i \le 1$; 2) $\sum_i p_i(x) = 1$.
- **§5.3 Mean, Variance, and Standard Deviation for a Discrete Random Variable**
  - For discrete R.V. X with pmf $p(x)$, the **mean** or **expected value** of X is $\text{E}(X) = \mu = \mu_X = \sum_{\text{all } x} x \cdot p(x)$.
  - **Rules for Means**. For R.V.s X, Y and fixed numbers a, b and function $g(X)$, $\mu_{a+bX} = a + b\mu_X$. 2) $\mu_{X\pm Y} = \mu_X \pm \mu_Y$. 3) $\text{E}(g(X)) = \sum g(x_i)p_i$.
  - **Variance** of X is $\text{Var}(X) = \sigma^2 = \sigma_X^2 = \sum_{\text{all } x}(x - \mu)^2 \cdot p(x) = \text{E}[(X-\mu)^2] = \text{E}(X^2) - (\text{E}(X))^2 = \text{E}(X^2) - \mu^2$.
  - **Standard deviation** of X is $\sigma = \sigma_X = \sqrt{\sigma^2}$.
  - **Rules for Variance**. 1) $\sigma_{a+bX}^2 = b^2\sigma_X^2$. 2) If X, Y are independent R.V.s, then $\sigma_{X\pm Y}^2 = \sigma_X^2 + \sigma_Y^2$. $\sigma_{X\pm Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$. 3) If X, Y have correlation ρ, then $\sigma_{X\pm Y}^2 = \sigma_X^2 + \sigma_Y^2 \pm 2\rho\sigma_X\sigma_Y$.
- **§§5.4 & 5.5 Binomial and Poisson Distributions**
  - **Properties of binomial experiment (BInS).** 1) Only two possible outcomes. 2) Outcomes of trials are independent 3) Experiment consists of n fixed identical trials. 4) For each trial, the probability p of success remains the same.
  - **Binomial random variable** maps each outcome in a binomial experiment to a real number and is defined to be number of successes in n trials. $X \sim B(n, p)$. Probability of success denoted by p. $P(S) = p$, $P(F) = 1 - p$.
  - **Binomial probability distribution.** Suppose X is a binomial R.V. with n trials and probability of a success p: $X \sim B(n, p)$. Then $p(x) = P(X = x) = \binom{n}{x}p^x(1-p)^{n-x}$, $x = 0, 1, 2, \ldots, n$. $\binom{n}{x} = \frac{n!}{x!(n-x)!}$.
  - **Cumulative Probability Function. (cdf)**. $P(X \le x) = \sum_{k=0}^x P(X = k) = P(X = 0) + P(X = 1) + P(X = 2) + \cdots + P(X = x)$.
  - Mean and standard deviation of binomial distribution. If $X \sim B(n, p)$, then $\mu = np$, $\sigma^2 = np(1-p)$, $\sigma = \sqrt{np(1-p)}$.
  - **Poission R.V.** is a count of # times the specific event occurs during a given interval.
  - **Poisson Experiment**. 1) Probability that a particular event occurs in a given interval is the same for all units of equal size and is proportional to the size of the unit. 2) Number of events that occur in any interval

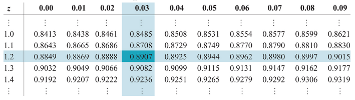
is independent of the number that occur in any other non-overlapping interval. 3) Probability that more than one event occurs in a unit of measure is negligible for very small units.
  - **Poisson Distribution**. For Poisson R.V. X with mean λ, $X \sim \text{Poisson}(\lambda)$, $p(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots, \mu_X = \sigma^2 = \lambda$, $\sigma_X = \sqrt{\lambda}$, $\lambda' = k\lambda$.

- **§6.1 Probability Distributions for a Continuous Random Variable**
  - **Probability distribution for a continuous random variable** X is given by smooth curve called density curve, or probability density function (pdf). Curve is defined so that probability that X takes on a value between a and b (for $a < b$) is the area under the curve between a and b.
  - Properties of pdf. 1) $f(x) \geq 0$. 2) $\int_{-\infty}^{\infty} f(x)\,dx = 1$.
  - Mean of a random variable. $\text{E}((g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\,dx$
  - **Variance of a Random Variable**. $\text{Var}(X) = \text{E}[(X-\mu)^2] = \sum(x-\mu_X)^2 \cdot p(x) = \int_{-\infty}^{\infty}(x-\mu_x)^2 f(x)\,dx$.
  - **Cumulative distribution function**. $F(x) = P(X \leq x) = P(X < x) = \int_{-\infty}^{x} f(s)\,ds$.
  - **Percentiles**. Let p be a number between 0 and 1. The 100pth percentile is defined by $p = \int_{-\infty}^{x} f(s)ds = F(x)$. Median of a pdf is the equal-areas point $p = 0.5 = \int_{-\infty}^{\tilde{\mu}} f(x)\,dx = F(\tilde{\mu})$.

- **§6.2 The Normal Distribution & §6.5 The Normal Approximation to the Binomial Distribution**
  - **Normal Distribution**. $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where $-\infty < \mu < \infty, \sigma > 0$. $X \sim N(\mu, \sigma^2)$.
  - Normal density curve is symmetric, bell-shaped, unimodal. μ is median. σ is spread. σ large $\implies$ curve is wider. σ small $\implies$ curve is narrower.
  - Graph of normal distribution. Inflection points at one standard deviation from mean in both directions.



  - **Example (z-table for $P(Z \leq 1.23)$).** Round to two decimal places value to look up.

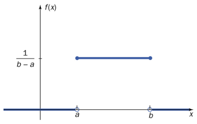


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

  - NOTE: Values for $P(Z > z)$ cannot be read directly from table. Use $P(Z > z) = 1 - P(Z \leq z)$. Also, $P(Z \leq z) = P(Z \geq -z)$. $P(z_1 < Z < z_2) = P(Z < z_2) - P(Z < z_1)$.
  - **Nonstandard normal distributions.** $z = \frac{x-\mu}{\sigma} \iff x = \mu + \sigma z$.
  - Procedure for normal distribution problems. 1) Sketch situation and shade area. 2) Standardize X to state problem in terms of Z. 3) Use Table III to find area to the left of z. 4) Calculate final answer. 5) Write conclusion in context of problem.
  - **Example**. A particular rash ... length of time that the rash will last is normally distributed with mean 6 days and S.D. 1.5 days. What interval symmetrically placed about mean will capture 95% of times for students' rashes to have lasted? **Soln.** $P(X < 6 + b) = 0.975$. $P(Z < \frac{6+b-6}{1.5}) = P(Z < \frac{b}{1.5}) = 0.975$. $\frac{b}{1.5} = 1.96 \implies b = 2.94$. $(6 - 2.94, 6 + 2.94) = (3.06, 8.94)$.
  - Use normal approximation when both $np \geq 10, n(1-p) \geq 10$, and when question asks for an interval. Continuity correction: $P(X \leq b) \approx P(X < b + 0.5)$.

- **§6.3 Checking the Normality Assumption**.
  - Methods for Checking Normality. 1) Graphs. 2) Backward Empirical Rule. 3) $\frac{\text{IQR}}{S}$. 4) Normal probability plot.
  - **Procedure: Normal Quantile Plot.** 1) Arrange data smallest to largest. 2) Record corresponding percentiles (quantiles). 3) Find z value corresponding to the quantile calculated in part 2. Plot original data points (from 1) vs. the z values (from 3).

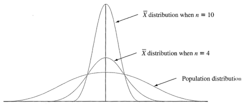
- **§6.4 The Exponential Distribution (and Uniform Distribution)**.
  - **Uniform Distribution.** Probability density is distributed evenly between two points. $\int_a^b c\,dx = c(b-a) = 1 \iff c = \frac{1}{b-a}$. $f(x) = \frac{1}{b-a}$ for $a < x < b$ and 0 elsewhere. $\text{E}(X) = \frac{a+b}{2}$. $\sigma_X = \frac{b-a}{\sqrt{12}}$.



  - **Exponential Distribution.** Amount of time until some specific event occurs. $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ and 0 elsewhere. $F(x) = 0, x < 0; 1 - e^{-\lambda x}, x \geq 0$. $\text{E}(X) = \sigma_X = \frac{1}{\lambda}$. $\text{Var}(X) = \frac{1}{\lambda^2}$.
  - Gamma distribution is generalization of exponential function and is used in probability theory, theoretical statistics, actuarial science, operations research, and engineering.
  - Beta distribution is defined on an interval – standard on $[0, 1]$. Used in modeling proportions, percentages, probabilities. Uniform distribution is member of this family.
  - Weibull (exponential) is used in lifetimes. lognormal is log. of normal distribution, used in products of distributions. Cauchy is symmetrical with long, straggly tails.

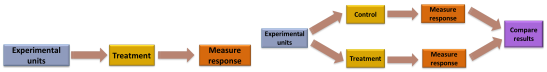
C7 SAMPLING DISTRIBUTIONS

- **§§7.1-7.2 Statistics, Parameters, Sampling Distribution of a Sample Mean**
  - **Parameter**: numerical descriptive measure of a population (Greek); **Statistic**: any quantity computed from values in a sample (Latin).
  - **Sampling distribution** of a statistic is the probability distribution of the statistic, distribution of values taken by the statistic in all possible samples of the same size from the same population. **Population distribution** of variable is the distribution of values of the variable among all individuals in the population.
  - μ_X is mean of population. σ_X is S.D. of population. $\mu_{\overline{X}}$ is mean of sampling distribution. $\sigma_{\overline{X}}$ is S.D. of sampling distribution.
  - **iid** means independent with identical distributions.
  - $\mu_{\overline{X}} = \mu_X$. $\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$. S.D. of sampling distribution decreases as number of samples increases.

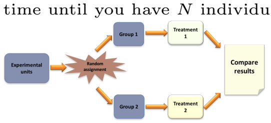


  - If population $X \sim N(\mu, \sigma^2)$, then sampling distribution $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.
  - Let $\overline{X}$ be the mean of observations in a random sample of size n drawn from a population with mean μ and finite variance σ^2. With n large enough, $\overline{X} \dot{\sim} N(\mu, \frac{\sigma^2}{n})$
  - Any linear combination of independent Normal R.V.s also is normal. The distribution of a sum or average of many small random quantities is close to Normal whether or not it is independent. C.L.T. applies also to discrete R.V.s.

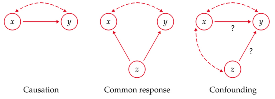
- **§1.3 Producing Data**
  - **Anecdotal data** represent individual cases that often come to our attention because they are striking in some way. **Available data** are data that were produced in the past for some other purpose but that may help answer a present question inexpensively; sources include Internet and library.
  - **Experimental study**: we investigate the effects of certain conditions on individuals or objects in the sample. **Observational study**: we observe the response for a specific variable for each individual or object.
  - **Experimental units**: the objects in which we are interested (human = subject). **Treatment / factor**: the specific experimental conditions we apply to the units. **Level** is the number of different values of the factor. **Outcome / response**: result. **Statistically significant**: response is larger than would be expected by chance.
  - CONTROL: compare two or more treatments.



  - **Biased**: systematically favors certain outcomes.
  - RANDOMIZE: use chance to assign experimental units to treatments. **Completely randomized design**: treatments are assigned to all experimental units completely by chance. Procedure: 1) Label each of N individuals. 2) Put N numbers into a hat. 3) Draw numbers one at a time until you have N individuals.



  - REPLICATION: use enough experimental units in each group to reduce chance variation in results.
  - Cautions: bias, generalization.
  - **Matched pair design**: when each experimental unit is matched with another one. **Block**: group of experimental units that are similar. **Block design**: random assignment of experimental units to treatments is carried out within each block. Rule: Control what you can control; block what you can't control; and randomize to create comparable groups.
  - **Probability sample**: method s.t. each sample is chosen by chance.
  - **Simple random sample (SRS)**: of size n is a sample selected s.t. every possible sample of size n has same chance of being selected. Procedure: 1) Label every object from 1 to n. 2) Generate random numbers to select objects. 3) Done when n different objects are selected.
  - **Stratified random sample**: 1) Divide population into groups called strata. 2) Choose SRS for each group.
  - **Response bias**: **Convenience sample**: sample is easy to obtain. **Undercoverage**: groups in population are left out of sample or are not represented by enough objects. **Nonresponse**.
  - Bias is to accuracy as variability is to precision. $\mathbb{E}(t) = \theta$, $\mathbb{E}(\overline{x}) = \mu$.
  - To reduce bias, use random sampling. To reduce variability, use a larger sample.
  - **Statistical inference**: Sample has to be representative of population. The experiment has to be performed such that you can obtain the data in which you are interested. Perform the correct analysis.
  - **Lurking variable**: variable that is not among the explanatory or response variables in a study but that may influence variables in the study. **Confounding** occurs when two variables are associated such that their effects on a response variable cannot be distinguished from each other.



Causation Common response Confounding

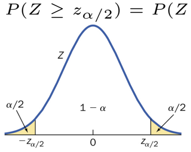
  - Needed for causation: Association is strong. Association is consistent. Connection happens in repeated trials. Same connection happens under varying conditions. Alleged cause precedes effect. Alleged cause is plausible.
  - **Simpson's Paradox**: an association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group.

C8 CONFIDENCE INTERVALS BASED ON A SINGLE SAMPLE

- **§8.1 Point Estimation**
  - **Point estimate** of a population parameter θ is a single number computed from a sample which serves as a best guess for the parameter.
  - **Estimator** is a statistic of interest and is a random variable. Has a distribution, mean, variance, and standard deviation. **Estimate** is a specific value of an estimator.
  - A statistic $\hat{\theta}$ is an **unbiased estimator** of a population parameter θ if $\mathbb{E}(\hat{\theta}) = \theta$. Otherwise, biased estimator. Among all estimators of θ

that are unbiased, choose the one that has minimum variance. This $\hat{\theta}$ is called **minimum variance unbiased estimator (MVUE)** of θ.

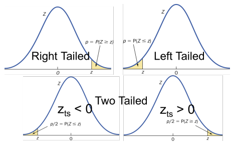
- **§8.2 A Confidence Interval for a Population Mean Where σ Is Known**
  - Assumptions for inference. 1) We have an S.R.S. from the population of interest. The variable we mean has a Normal distribution (or approx. Normal distribution) with mean μ and S.D. σ. 3) We don't know μ.
  - **Confidence interval** (estimate $\pm$ margin of error): for a population parameter is an interval of values constructed so that with a specified degree of confidence, the value of the population parameter lies in this interval. C, **confidence coefficient**, is probability that C.I. encloses the population parameter in repeated samplings. **Confidence level** is confidence coefficient expressed as percentage.
  - $z_{\alpha/2}$ is value on measurement axis in a standard normal distribution s.t. $P(Z \geq z_{\alpha/2}) = P(Z \leq -z_{\alpha/2}) = \alpha/2$ and $P(Z \leq z_{\alpha/2}) = 1 - \alpha/2$.

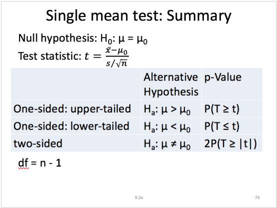


  - C.I. is given by $\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Margin of error is given by $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. $n = (z_{\alpha/2}\sigma/\text{ME})^2$
  - HIGHER CONFIDENCE LEVEL IMPLIES LOWER PRECISION.
  - To increase precision of confidence level: Lower M.E., Lower C, reduce σ, increase n.
  - Upper confidence bound: $\mu < \overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$. Lower confidence bound: $\mu > \overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}$
  - INCREASE sample size gives REDUCED width.
- **§8.3 Inference for the Mean of a Population**
  - $t_{\alpha,\nu}$ is critical value for a t distribution with ν degrees of freedom. ROUND DOWN for degrees of freedom.
  - $P(T \geq t_{\alpha/2,\nu}) = P(T \leq -t_{\alpha/2,\nu}) = \alpha/2$
  - Confidence interval: $\overline{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$. Upper confidence bound: $\mu < \overline{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}$. Lower confidence bound: $\mu > \overline{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}}$.
  - A statistical value or procedure is robust if the calculations required are insensitive to violations of the conditions. t procedure is robust against normality. $n < 15 \implies$ population dist. close to normal. $15 < n < 40 \implies$ mild skewedness is acceptable. $n > 40 \implies$ procedure usually valid.

C9 HYPOTHESIS TESTS BASED ON A SINGLE SAMPLE

- **§9.1 The Parts of a Hypothesis Test**
  - **Hypothesis**: a declaration, or claim, in the form of a mathematical statement, about the value of a specific population parameter (or about the values of several population characteristics). **Hypothesis test**: a formal procedure form comparing observed data with a claim (hypothesis) whose truth we want to assess.
  - Parts of hypothesis test: 1) Claim assumed to be true, 2) alternative claim, 3) How to test claim, 4) what to use to make decision.
  - H_0 is assumed to be true. H_a is contradictory to H_0.
  - H_0 always has an equal sign. $H_a = \mu > \mu_0$ is an upper tail. $H_a = \mu < \mu_0$ is a lower tail. $H_a = \mu \neq \mu_0$ is a two-tailed.
  - **Test statistic** calculated from sample data measures how far data diverge from what we would expect if the null hypothesis H_0 were true.
  - **p-value** for a hypothesis test is the smallest significance level for which the null hypothesis H_0 can be rejected.
- **§9.2 Hypothesis Test Errors and Powers**
  - **Type I error**: Reject H_0 when H_0 is true. $P(\text{Type I error}) = \alpha$. **Type II error**: fail to reject H_0 when H_0 is false. $P(\text{Type II error}) = \beta$. Power $= 1 - \beta$.
  - α measures the strength of the sample evidence against H_0. Power measures sensitivity (true negative) of test.
  - α, μ_a, σ, n all affect power.
  - β decreases, then power increases. Increase α, β decreases. Real mean translated to right, both α and β decrease. Decrease σ, decrease widths, decrease both α and β; same for increase n.
  - To find $\beta(c)$. Find z_α and solve for $\overline{x}$. $\beta(b) = P_{H_a}(\overline{X} \geq\leq b)$. Normalize.
- **§§9.3 & 9.4 Hypothesis Tests Concerning a Population Mean When σ Is Known**
  - **Test statistic** calculated from sample data measures how far the data divere from what we would expect if H_0 true.
  - $z_{ts} = \frac{\text{estimate}-\text{hypothesized value}}{\text{S.D. of estimate}} = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$. TS large means data are not consistent w/ H_0. TS small show that data is consistent with H_0.





Single mean test: Summary

Null hypothesis: H$_0$: $\mu = \mu_0$

Test statistic: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$

	Alternative Hypothesis	p-Value
One-sided: upper-tailed	H$_a$: $\mu > \mu_0$	P(T ≥ t)
One-sided: lower-tailed	H$_a$: $\mu < \mu_0$	P(T ≤ t)
two-sided	H$_a$: $\mu \neq \mu_0$	2P(T ≥ |t|)

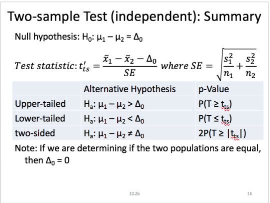
df = n - 1

9.5a 73

  - Probability assuming H_0 true that statistic would take a value as extreme or more extreme than the one actually observed is p-value. Smaller p, stronger evidence against H_0. Smallest significance level for which H_0 can be rejected.
  - α measures strength of sample evidence against H_0. p-value is smaller than α, data are **statistically significant at level** α. Quantity α is called **significance level**.
  - If $p \leq \alpha$, reject H_0, conclude H_a in context. $p > \alpha$, fail to reject H_0, cannot conclude H_a in context.
  - Procedure for Hypothesis Testing: 1) Identify of parameters of interest and describe them in context or problems. 2) State hypothesis. 3) Calculate appropriate test statistic and find p-value if appropriate. 4) Make the decision (with reason) and state the conclusion in problem context. The data (does/might) (not) give (strong) support (p-value) to claim that (H_a).
  - Non-significant answer means that experimental data is consistent with null.
- **§§9.5 Hypothesis Tests Concerning a Population Mean When σ Is Unknown**
  - z-test when σ is known. t-test when σ is unknown.

C10 C.I. and H.T. Based on Two Samples or Treatments

- **§§10.1-10.2 Comparing Two Population Means Using Independent Samples**
  - Two samples are **independent** if the process of selecting individuals or objects in sample 1 has no effect on, or no relation to, the selection of individuals or objects in sample 2. A **paired** data set is the result of matching each individual or object in sample 1 with a similar individual or object in sample 2.
  - $\mathbb{E}(\overline{X}_1) - \mathbb{E}(\overline{X}_2) = \mu_1 - \mu_2$. $\text{Var}(\overline{X}_1 - \overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.
  - **Two-sample independent Z Hypothesis Test**. 1) Define populations. 2) $H_0 : \mu_1 - \mu_2 = \Delta_0$. 3) $z_{ts} = \frac{(\overline{x}_1 - \overline{x}_2) - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$.
  - The $100(1-\alpha)\%$ C.I. for $\mu_1 - \mu_2$ is $(\overline{x}_1 - \overline{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.
  - If two pop. variances are unknown but assumed to be equal, we call this **pooling**.
  - **Two-sample independent t Hypothesis Test**: $t_{ts} = \frac{(\overline{x}_1 - \overline{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$.
  - Satterthwaite Approximation. $$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$
  -



Two-sample Test (independent): Summary

Null hypothesis: H$_0$: $\mu_1 - \mu_2 = \Delta_0$

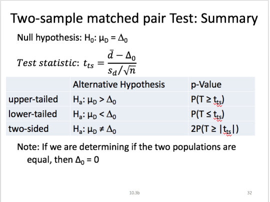
Test statistic: $t'_{ts} = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{SE}$ *where* $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

	Alternative Hypothesis	p-Value
Upper-tailed	H$_a$: $\mu_1 - \mu_2 > \Delta_0$	P(T ≥ t$_{ts}$)
Lower-tailed	H$_a$: $\mu_1 - \mu_2 < \Delta_0$	P(T ≤ t$_{ts}$)
two-sided	H$_a$: $\mu_1 - \mu_2 \neq \Delta_0$	2P(T ≥ |t$_{ts}$|)

Note: If we are determining if the two populations are equal, then $\Delta_0 = 0$

10.2b 16

  - C.I. $(\overline{x}_1 - \overline{x}_2) \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
  - t-procedure is very robust against normality.
- **§10.3 Paired Data**
  - Find difference between responses within each pair. $df = n - 1$. $\mathbb{E}(\overline{D}) = \mu_1 - \mu_2$. $\text{Var}(\overline{D}) = \sigma_D/\sqrt{n}$.
  - $SE = s_D/\sqrt{n}$. Confidence interval: $\overline{d} \pm t_{\alpha/2,n-1}\frac{s_D}{\sqrt{n}}$.
  -



Two-sample matched pair Test: Summary

Null hypothesis: H$_0$: $\mu_D = \Delta_0$

Test statistic: $t_{ts} = \frac{\bar{d} - \Delta_0}{s_d/\sqrt{n}}$

	Alternative Hypothesis	p-Value
upper-tailed	H$_a$: $\mu_D > \Delta_0$	P(T ≥ t$_{ts}$)
lower-tailed	H$_a$: $\mu_D < \Delta_0$	P(T ≤ t$_{ts}$)
two-sided	H$_a$: $\mu_D \neq \Delta_0$	2P(T ≥ |t$_{ts}$|)

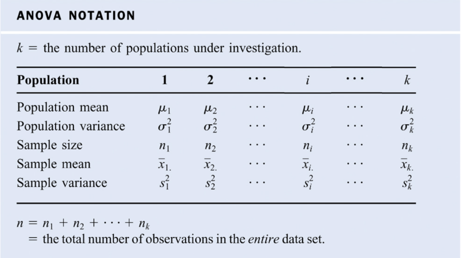
Note: If we are determining if the two populations are equal, then $\Delta_0 = 0$

10.3b 32

  - If there is great heterogeneity between experimental units and a large correlation within experimental units then a paired experiment is preferred. If the experimental units are relatively homogeneous and the correlation within pairs is not large, then unpaired experiments should be used.

C11 Analysis of Variance (ANOVA)

- **§11.1 One-way ANOVA**
  - **Factor** is what differentiates the population. **Level** or **group** is the number of different populations k.
  - **One-way ANOVA** is used for situations in which there is only one factor or only one way to classify the populations of interest. **Two-way ANOVA** is used to analyze the effect of two factors.
  -



ANOVA NOTATION

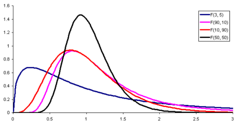
k = the number of populations under investigation.

Population	**1**	**2**	$\cdots$	i	$\cdots$	k
Population mean	μ_1	μ_2	$\cdots$	μ_i	$\cdots$	μ_k
Population variance	σ_1^2	σ_2^2	$\cdots$	σ_i^2	$\cdots$	σ_k^2
Sample size	n_1	n_2	$\cdots$	n_i	$\cdots$	n_k
Sample mean	$\bar{x}_{1.}$	$\bar{x}_{2.}$	$\cdots$	$\bar{x}_{i.}$	$\cdots$	$\bar{x}_{k.}$
Sample variance	s_1^2	s_2^2	$\cdots$	s_i^2	$\cdots$	s_k^2

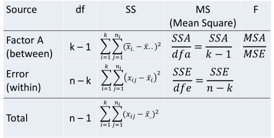
$n = n_1 + n_2 + \cdots + n_k$
= the total number of observations in the *entire* data set.

  - $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$. $H_a : \mu_i \neq \mu_j$ for some $i \neq j$.
  - Assumptions for ANOVA: 1) We have k independent SRSs, one from each population. We measure the same response variable for each sample. 2) The ith population has a Normal distribution with unknown mean μ_i. Check for Normality with QQ plot and histogram. 3) All the populations have the same variance σ^2, whose value is unknown. $\frac{s_{\max}}{s_{\min}} \leq 2$ is how to check for constant standard deviation.
  - Terminology for ANOVA: $x_{ij}; i = 1, 2, \ldots, k; j = 1, 2, \ldots, n_i$. i is group or level, k total levels. j is which object we discuss in group i. Number of objects in group i is n_i. μ_i.
  - Model: $X_{ij} = \mu_i + \epsilon_{ij}$. $\epsilon_{ij} \sim^{\text{iid}} N(0, \sigma^2)$ is error term. DATA = FIT + RESIDUAL.
  - $\overline{x}_{i.} = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \approx \mu_i$. $\overline{x}_{..} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}}{n}$.
  - $s_i^2 = \frac{\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2}{n_i - 1} = \frac{\text{SS}}{df}$.
  - test statistic $= \frac{\text{between}-\text{sample variation}}{\text{within}-\text{sample variation}}$. H_0 true gives small fraction. H_0 false gives large fraction.

- **Analysis of variation** compares the variation due to specific sources with the variation among individuals who should be similar. In particular, ANOVA tests whether several populations have the same means by comparing how far apart the sample means are with how much variation there is within a sample.
- Var $= \frac{\text{SS}}{n-1} = \frac{\text{SS}}{df} =$ MS, where SS is sum of squares, and MS is mean square.
- SSM (SS for model) or SSG (SS for groups) or SSA (SS for factor A): between samples. SSA $= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{x}_{i\cdot} - \overline{x}_{i\cdot\cdot})^2 = \sum_{i=1}^{k} n_i(\overline{x}_{i\cdot} - \overline{x}_{i\cdot\cdot})^2$. $dfa = k-1$. $MSA = \frac{SSA}{dfa}$.
- SSE (SS for error) or SSR (SS for residuals): SSE $= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_{i\cdot})^2 = \sum_{i=1}^{k}(n_i-1)s_i^2$. $dfe = n-k$. $MSE = \frac{SSE}{dfe}$.
- SST (SS for total) $SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x})^2$. $dft = n-1$. $SST = SSE + SSA$. $dft = dfe + dfa$.



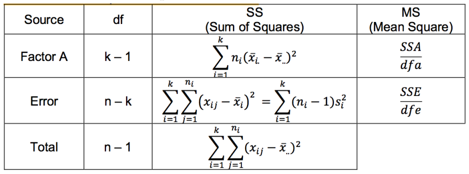
- **F distribution**.
- $F_{ts} = MSA/MSE; df_1 = dfa; df_2 = dfe$.
-



Source	df	SS	MS (Mean Square)	F
Factor A (between)	k – 1	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_{i.} - \bar{x}_{..})^2$	$\frac{SSA}{dfa} = \frac{SSA}{k-1}$	$\frac{MSA}{MSE}$
Error (within)	n – k	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$	$\frac{SSE}{dfe} = \frac{SSE}{n-k}$	
Total	n – 1	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{..})^2$		

- p-value: $P(F \geq F_{ts})$ has a $F_{\alpha,dfa,dfe}$ dist.
- Example: 1) List μs. 2) List hypotheses. 3) Calculate dfa, dfe, dft. Calculate $F_{ts} = MSA/MSE$. Calculate $df_1 = dfa$ and $df_2 = dfe$. 4) Decision and conclusion.
- Advantages of t distribution: Don't have to pool. F test has to be two-tailed. Only two groups, use t test. More than 2 groups, use ANOVA.

- **§11.2 Isolating Differences**
  - To determine which mean is different: 1) Graphics. 2) Multiple comparisons
  - **Simultaneous Confidence Intervals**. $\overline{x}_{i\cdot} - \overline{x}_{j\cdot} \pm t_{\text{column, df}} \cdot SE$. $\overline{x}_{i\cdot} - \overline{x}_{j\cdot} \pm t^{**}_{\text{column, df}} \cdot \sqrt{MSE(1/n_i + 1/n_j)}$. $\overline{x}_{i\cdot} - \overline{x}_{j\cdot} \pm t^{**}_{\text{column, df}} \cdot \sqrt{2MSE/n_i}$.
  - How many tests? $\binom{k}{2} = k(k-1)/2 = c$.
  - **Tukey method**. $t^{**}_{\text{column, df}} = \frac{Q_{\alpha,k,n-k}}{\sqrt{2}}$.
  - Procedure: 1) Perform the ANOVA test; only continue if results are statistically significant. 2) Select a family significant level α. 3) Select multiple comparison methodology. 4) Calculate t^{**}. 5) Calculate all confidence intervals required by procedure. 6) Determine which ones are statistically significant. 7) Visually display results. Order means. Draw a line under values that are not significantly different (i.e., contains 0). 8) Write conclusion in context of problem.
  - ANOVA table.

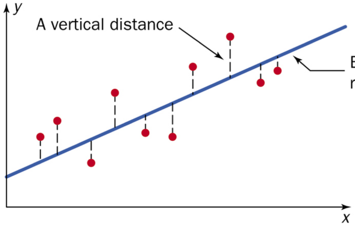


Source	df	SS (Sum of Squares)	MS (Mean Square)
Factor A	k – 1	$\sum_{i=1}^{k} n_i(\bar{x}_{i.} - \bar{x}_{..})^2$	$\frac{SSA}{dfa}$
Error	n – k	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^{k}(n_i - 1)s_i^2$	$\frac{SSE}{dfe}$
Total	n – 1	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{..})^2$	

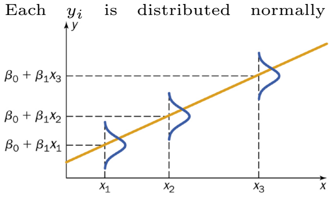
  - Use Tukey's method when making all pairwise comparisons.
  - Approximate binomial distribution: $\mu = np, \sigma = \sqrt{np(1-p)}$. $np \geq 10, n(1-p) \geq 10$.

C12 CORRELATION AND LINEAR REGRESSION

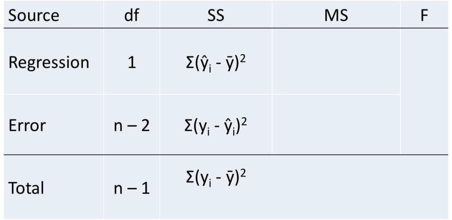
- **§12.1 Simple Linear Regression**
  - **Association**. Two variables are associated if knowing the values of one of the variables tells you something about the values of the other variables. NOT same as causation.
  - **Response variable** Y is outcome of study. **Explanatory variable** X expalines or causes changes in the response variable. $Y = g(X)$.
  - Procedure for scatterplots. 1) Place explanatory variable on X-axis and response variable on Y-axis. 2) Label and scale axes. 3) Plot (X, Y) pairs.
  - Interpreting a scatterplot. Form – linear, curved, clusters, no pattern. Direction – positive, negative, horizontal. Strength – how close the points are to the line. Outliers.
  - A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes.
  - $Y = \beta_0 + \beta_1 X + \epsilon$.
  - Notation: n independent observations, x_i are the explanatory observations, y_i are the observed response variable observations. We have n ordered pairs (x_i, y_i).
  - Let (x_i, y_i) be pairs of observations. We assume that there exist constants β_0 and β_1 such that $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where $\epsilon_i \sim^{\text{iid}} N(0, \sigma^2)$. Then $\mathbb{E}(Y_i) = \beta_0 + \beta_1 X_i$. $Y_i = Y \mid X_i$. $\text{Var}(Y_i) = \text{Var}(\epsilon_i) = \sigma^2$.



  -
  - Assumptions for linear regression. 1) SRS with observations independent of each other. 2) The relationship in the population is linear. The response variable is normally distributed around the population regression line. The S.D. of the response is constant.
  - Each y_i is distributed normally around the regression line.



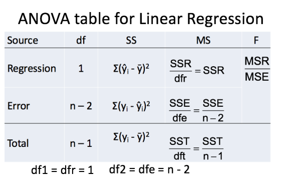
  - $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = b_0 + b_1 x$. $\hat{\beta}_1 = b_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{S_{XY}}{S_{XX}}$. $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$. $b_0 = \overline{y} - b_1\overline{x}$. $y^* = b_0 + b_1 x^*$.
  - **Linear regression - variance**. $\text{e}_i = y_i\hat{y}_i$. $s^2 = MSE$.



Source	df	SS	MS	F
Regression	1	$\Sigma(\hat{y}_i - \bar{y})^2$		
Error	n – 2	$\Sigma(y_i - \hat{y}_i)^2$		
Total	n – 1	$\Sigma(y_i - \bar{y})^2$		

  -
  - Horizontal line implies no association.
  - $SST = S_{YY} = \sum(y_i - \overline{y})^2$. $dft = n - 1$
  - $SSR = \sum(\hat{y}_i - \overline{y})^2 = b_1 S_{XY}$. $dfr = 1$. $SSE = SST - SSR$.
  - $r^2 = \frac{SSR}{SST}$

- **§12.2 Hypothesis Tests**
  - H_0 : there is no association between X and Y. H_a : there is an association between X and Y.

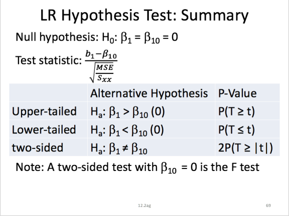


ANOVA table for Linear Regression

Source	df	SS	MS	F
Regression	1	$\Sigma(\hat{y}_i - \bar{y})^2$	$\frac{SSR}{dfr} = SSR$	$\frac{MSR}{MSE}$
Error	n – 2	$\Sigma(y_i - \hat{y}_i)^2$	$\frac{SSE}{dfe} = \frac{SSE}{n-2}$	
Total	n – 1	$\Sigma(y_i - \bar{y})^2$	$\frac{SST}{dft} = \frac{SST}{n-1}$	

df1 = dfr = 1   df2 = dfe = n - 2

  - F **test statistic**. $F = \frac{MSR}{MSE}$
  - $b_1 = \sum a_i y_i$. $\sigma_{b_1} = \frac{\sigma}{\sqrt{S_{xx}}}$.
  - $SE_{b_1} = \sqrt{\frac{MSE}{S_{xx}}}$.
  - **Confidence interval for** β_1: $b_1 \pm t_{\alpha/2,df} SE_{b_1} = b_1 \pm t_{\alpha/2,n-2}\sqrt{\frac{MSE}{S_{xx}}}$.



LR Hypothesis Test: Summary

Null hypothesis: H_0: $\beta_1 = \beta_{10} = 0$

Test statistic: $\frac{b_1 - \beta_{10}}{\sqrt{\frac{MSE}{S_{xx}}}}$

	Alternative Hypothesis	P-Value
Upper-tailed	H_a: $\beta_1 > \beta_{10}$ (0)	P(T ≥ t)
Lower-tailed	H_a: $\beta_1 < \beta_{10}$ (0)	P(T ≤ t)
two-sided	H_a: $\beta_1 \neq \beta_{10}$	2P(T ≥ |t|)

Note: A two-sided test with β_{10} = 0 is the F test

12.2ag 69

  -
  - **Sample Correlation**: r is measure of the strength of a linear relationship between two continuous.
  - Simple linear regression: $r^2 = R^2$. $r > 0$ implies positive association. $r < 0$ negative association.
  - H_0 is no association between X and Y. H_a there is an association between X and Y. $F_{ts} = MSR/MSE$. $p = P(F > F_{ts})$. $df_1 = dfr = 1, df_2 = dfe = n - 2$.
  - $r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$.

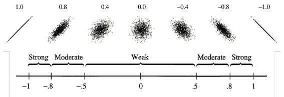


Figure 3.6 Describing the strength of relationship

  - Properties

- **§12.4 Regression Diagnostics**
  - Assumptions for Linear Regression. 1) SRS with observations independent of each other. 2) The relationship is linear in the population. 3) The standard deviation of the response is constant. 4) The response y is normally distributed around the population regression line.
  - Residuals: observed − predicted. Easier to look at horizontal line. Scale is larger. Should look like horizontal band around 0 with no discernible pattern.

- **§12.3 Inferences Concerning the Mean Value and Observed Value of Y for $x = x*$.**
  - $y^* = b_0 + b_1 x^*$. $\mu^* = \beta_0 + \beta_1 x^* + \epsilon$. $\mathbb{E}(\mu^*) = \hat{\mu}^* = b_0 + b_1 x^*$. $\hat{\mu}^*$ is an unbiased estimator of μ^*. $SE_{\hat{\mu}^*} = \sqrt{MSE[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}}]}$. Confidence interval given by $\hat{\mu}^* \pm t_{\alpha/2,n-2} SE_{\hat{\mu}^*}$
  - $SE_{\hat{y}^*} = \sqrt{MSE[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}}]}$ for PREDICTION INTERVAL. $\hat{y}^* \pm t_{\alpha/2,n-2} SE_{\hat{y}^*}$