

Formulas for AP Statistics

I. Descriptive Statistics

$$\bar{x} = \frac{1}{n} \sum x_i = \frac{\sum x_i}{n}$$

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

$$\hat{y} = a + bx$$

$$\bar{y} = a + b\bar{x}$$

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

$$b = r \frac{s_y}{s_x}$$

II. Probability and Distributions

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Probability Distribution	Mean	Standard Deviation
Discrete random variable, X	$\mu_X = E(X) = \sum x_i P(x_i)$	$\sigma_X = \sqrt{\sum (x_i - \mu_X)^2 P(x_i)}$
If X has a binomial distribution with parameters n and p , then: $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ where $x = 0, 1, 2, 3, \dots, n$	$\mu_X = np$	$\sigma_X = \sqrt{np(1-p)}$
If X has a geometric distribution with parameter p , then: $P(X = x) = (1-p)^{x-1} p$ where $x = 1, 2, 3, \dots$	$\mu_X = \frac{1}{p}$	$\sigma_X = \frac{\sqrt{1-p}}{p}$

III. Sampling Distributions and Inferential Statistics

Standardized test statistic: $\frac{\text{statistic} - \text{parameter}}{\text{standard error of the statistic}}$
Confidence interval: $\text{statistic} \pm (\text{critical value})(\text{standard error of statistic})$

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

III. Sampling Distributions and Inferential Statistics (*continued*)

Sampling distributions for proportions:

Random Variable	Parameters of Sampling Distribution		Standard Error* of Sample Statistic
For one population: \hat{p}	$\mu_{\hat{p}} = p$	$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$	$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
For two populations: $\hat{p}_1 - \hat{p}_2$	$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$	$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ When $p_1 = p_2$ is assumed: $s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ where $\hat{p}_c = \frac{X_1 + X_2}{n_1 + n_2}$

Sampling distributions for means:

Random Variable	Parameters of Sampling Distribution		Standard Error* of Sample Statistic
For one population: \bar{X}	$\mu_{\bar{X}} = \mu$	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$	$s_{\bar{X}} = \frac{s}{\sqrt{n}}$
For two populations: $\bar{X}_1 - \bar{X}_2$	$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$	$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Sampling distributions for simple linear regression:

Random Variable	Parameters of Sampling Distribution		Standard Error* of Sample Statistic
For slope: b	$\mu_b = \beta$	$\sigma_b = \frac{\sigma}{\sigma_x \sqrt{n}}$, where $\sigma_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$	$s_b = \frac{s}{s_x \sqrt{n-1}}$, where $s = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$ and $s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$

*Standard deviation is a measurement of variability from the theoretical population. Standard error is the estimate of the standard deviation. If the standard deviation of the statistic is assumed to be known, then the standard deviation should be used instead of the standard error.

Table entry for p and C is the point t^* with probability p lying above it and probability C lying between $-t^*$ and t^* .

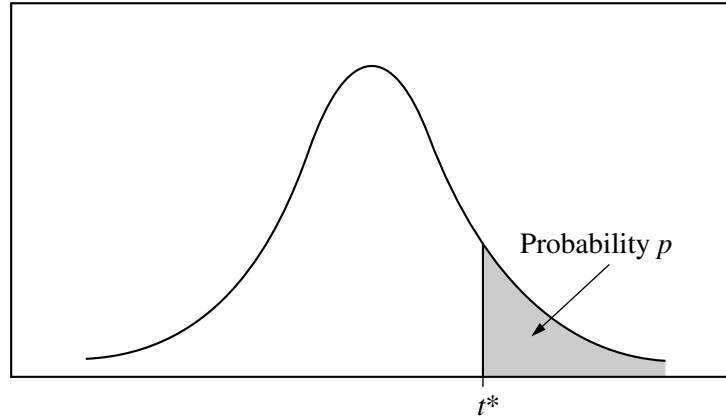
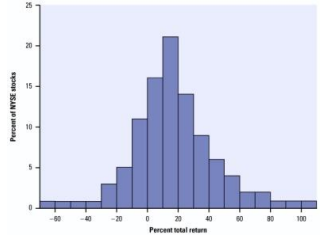
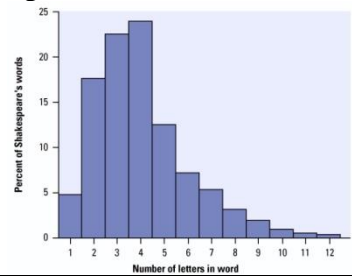
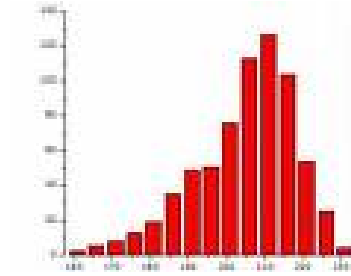
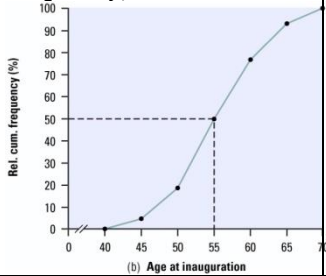
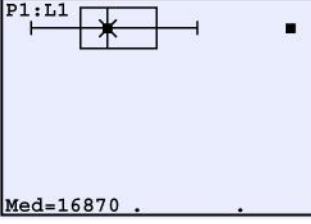
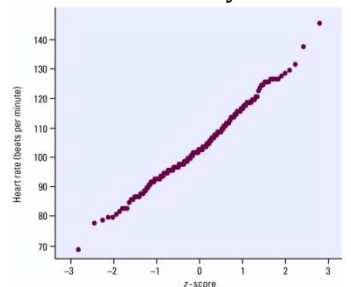
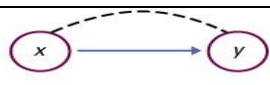
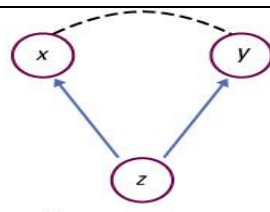
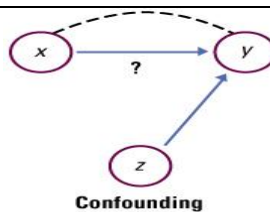


Table B t distribution critical values

df	Tail probability p											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
Confidence level C												

Important Concepts not on the AP Statistics Formula Sheet

Part I:

<p>IQR = $Q_3 - Q_1$ Test for an outlier: $1.5(IQR)$ above Q_3 or below Q_1 The calculator will run the test for you as long as you choose the boxplot with the outlier on it in STATPLOT</p>	<p>Linear transformation: Addition: affects center NOT spread adds to \bar{x}, M, Q_1, Q_3, IQR not σ Multiplication: affects both center and spread multiplies \bar{x}, M, Q_1, Q_3, IQR, σ</p>	<p>When describing data: describe center, spread, and shape. Give a 5 number summary or mean and standard deviation when necessary.</p>	<p>Histogram: fairly symmetrical unimodal</p> 
<p>skewed right</p> 	<p>Skewed left</p> 	<p>Ogive (cumulative frequency)</p> 	<p>Boxplot (with an outlier)</p> 
<p>Stem and leaf</p> <p>Treasury bills</p> <pre> 0 9 1 0 2 5 5 6 6 6 8 2 1 5 7 7 9 3 0 1 1 3 5 5 8 9 9 4 2 4 7 7 8 5 1 1 2 2 2 5 6 6 7 8 7 9 6 2 4 5 6 9 7 2 7 8 8 0 4 8 9 8 10 4 5 11 3 12 13 14 7 </pre> <p style="text-align: center;">(b)</p>	<p>Normal Probability Plot</p>  <p>The 80th percentile means that 80% of the data is below that observation.</p>	$z = \frac{x - \text{mean}}{\text{standard dev}}$ <p style="text-align: center;">or</p> $z = \frac{x - \mu}{\sigma}$ <p>HOW MANY STANDARD DEVIATIONS AN OBSERVATION IS FROM THE MEAN</p> <p>68-95-99.7 Rule for Normality $N(\mu, \sigma)$ $N(0, 1)$ Standard Normal</p>	<p>r: correlation coefficient, The strength of the linear relationship of data. Close to 1 or -1 is very close to linear</p> <p>r^2: coefficient of determination. How well the model fits the data. Close to 1 is a good fit. "Percent of variation in y described by the LSRL on x"</p>
<p>residual = $y - \hat{y}$</p> <p>residual = observed - predicted</p> <p>$y = a + bx$ Slope of LSRL(b): rate of change in y for every unit x</p> <p>y-intercept of LSRL(a): y when x = 0</p>	<p>Exponential Model: $y = ab^x$ take log of y</p> <p>Power Model: $y = ax^b$ take log of x and y</p>	<p>Explanatory variables explain changes in response variables. EV: x, independent RV: y, dependent</p>	<p>Lurking Variable: A variable that may influence the relationship between two variables. LV is not among the EV's</p>
<p>Confounding: two variables are confounded when the effects of an RV cannot be distinguished.</p>	 <p style="text-align: center;">Causation (a)</p>	 <p style="text-align: center;">Common response (b)</p>	 <p style="text-align: center;">Confounding (c)</p>

Regression in a Nutshell

Given a Set of Data:

Data:

NEA change (cal):	-94	-57	-29	135	143	151	245	355
Fat gain (kg):	4.2	3.0	3.7	2.7	3.2	3.6	2.4	1.3
NEA change (cal):	392	473	486	535	571	580	620	690
Fat gain (kg):	3.8	1.7	1.6	2.2	1.0	0.4	2.3	1.1

Enter Data into L₁ and L₂ and run **8:Linreg(a+bx)**

The regression equation is:

$$\text{predicted fat gain} = 3.5051 - 0.00344(\text{NEA})$$

y-intercept: Predicted fat gain is 3.5051 kilograms when NEA is zero.

slope: Predicted fat gain decreases by .00344 for every unit increase in NEA.

r: correlation coefficient

$$r = -0.778$$

Moderate, negative correlation between NEA and fat gain.

r²: coefficient of determination

$$r^2 = 0.606$$

60.6% of the variation in fat gained is explained by the Least Squares Regression line on NEA.

The linear model is a moderate/reasonable fit to the data. It is not strong.

The residual plot shows that the model is a reasonable fit; there is not a bend or curve, There is approximately the same amount of points above and below the line. There is No fan shape to the plot.

Predict the fat gain that corresponds to a NEA of 600.

$$\text{predicted fat gain} = 3.5051 - 0.00344(600)$$

$$\text{predicted fat gain} = 1.4411$$

Would you be willing to predict the fat gain of a person with NEA of 1000?

No, this is extrapolation, it is outside the range of our data set.

Residual:

observed y - predicted y

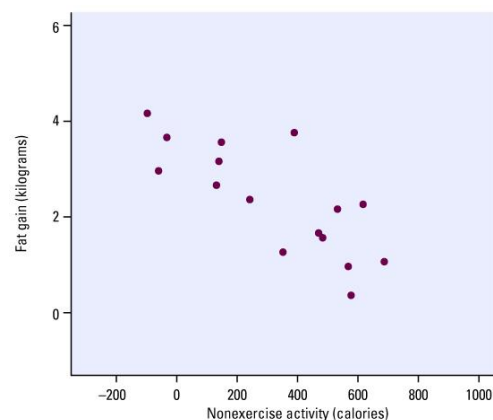
Find the residual for an NEA of 473

First find the predicted value of 473:

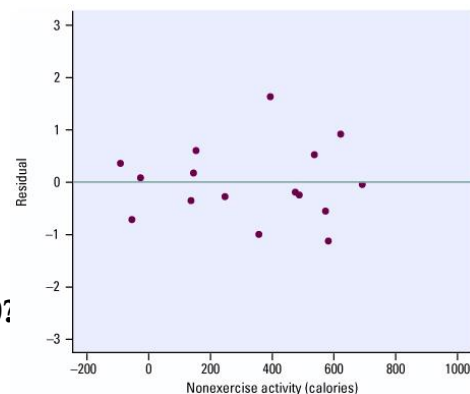
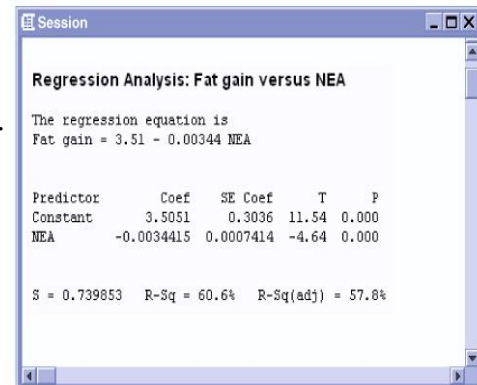
$$\text{predicted fat gain} = 3.5051 - 0.00344(473)$$

$$\text{predicted fat gain} = 1.87798$$

$$\text{observed} - \text{predicted} = 1.7 - 1.87798 = -0.17798$$



Minitab



Transforming Exponential Data $y = ab^x$

Take the log or ln of y.

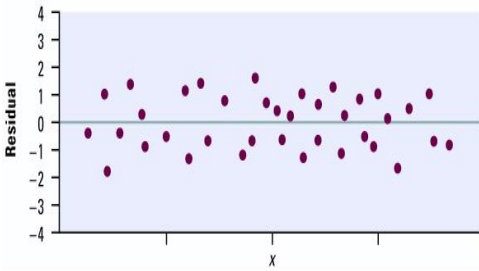
The new regression equation is:
 $\log(y) = a + bx$

Transforming Power Data $y = ax^b$

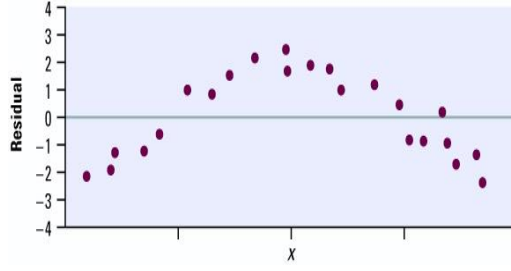
Take the log or ln of x and y.

The new regression equation is:
 $\log(y) = a + b \log(x)$

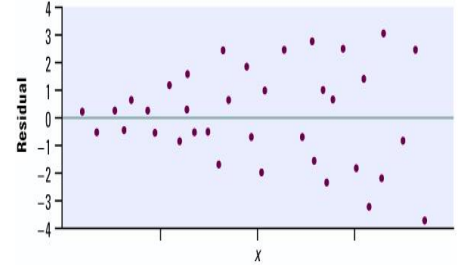
Residual Plot examples:



Linear mode is a Good Fit

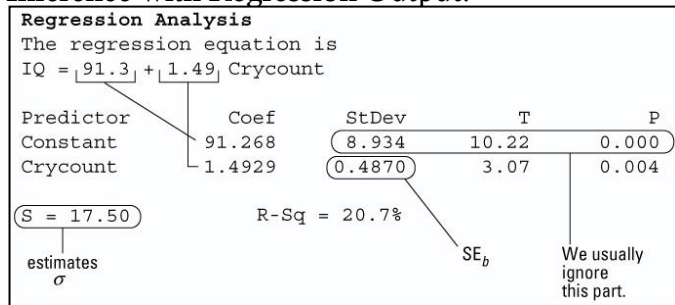


Curved Model would be a good fit



Fan shape loses accuracy as x increases

Inference with Regression Output:



Construct a 95% Confidence interval for the slope of the LSRL of IQ on cry count for the 20 babies in the study.

Formula: $df = n - 2 = 20 - 2 = 18$

$$b \pm t^* SE_b$$

$$1.4929 \pm (2.101)(0.4870)$$

$$1.4929 \pm 1.0232$$

$$(0.4697, 2.5161)$$

Find the t-test statistic and p-value for the effect cry count has on IQ.

From the regression analysis $t = 3.07$ and $p = 0.004$

Or

$$t = \frac{b}{SE_b} = \frac{1.4929}{0.4870} = 3.07$$

s = 17.50

This is the standard deviation of the residuals and is a measure of the average spread of the deviations from the LSRL.

Part II: Designing Experiments and Collecting Data:

Sampling Methods:

The Bad:

Voluntary sample. A voluntary sample is made up of people who decide for themselves to be in the survey.

Example: Online poll

Convenience sample. A convenience sample is made up of people who are easy to reach.

Example: interview people at the mall, or in the cafeteria because it is an easy place to reach people.

The Good:

Simple random sampling. Simple random sampling refers to a method in which all possible samples of n objects are equally likely to occur.

Example: assign a number 1-100 to all members of a population of size 100. One number is selected at a time from a list of random digits or using a random number generator. The first 10 selected without repeats are the sample.

Stratified sampling. With stratified sampling, the population is divided into groups, based on some characteristic. Then, within each group, a SRS is taken. In stratified sampling, the groups are called **strata**.

Example: For a national survey we divide the population into groups or strata, based on geography - north, east, south, and west. Then, within each stratum, we might randomly select survey respondents.

Cluster sampling. With cluster sampling, every member of the population is assigned to one, and only one, group. Each group is called a cluster. A sample of clusters is chosen using a SRS. Only individuals within sampled clusters are surveyed.

Example: Randomly choose high schools in the country and only survey people in those schools.

Difference between cluster sampling and stratified sampling. With stratified sampling, the sample includes subjects from each stratum. With cluster sampling the sample includes subjects only from sampled clusters.

Multistage sampling. With multistage sampling, we select a sample by using combinations of different sampling methods.

Example: Stage 1, use cluster sampling to choose clusters from a population. Then, in Stage 2, we use simple random sampling to select a subset of subjects from each chosen cluster for the final sample.

Systematic random sampling. With systematic random sampling, we create a list of every member of the population. From the list, we randomly select the first sample element from the first k subjects on the population list. Thereafter, we select every k th subject on the list.

Example: Select every 5th person on a list of the population.

Experimental Design:

A well-designed experiment includes design features that allow researchers to eliminate extraneous variables as an explanation for the observed relationship between the independent variable(s) and the dependent variable.

Experimental Unit or Subject: The individuals on which the experiment is done. If they are people then we call them subjects

Factor: The explanatory variables in the study

Level: The degree or value of each factor.

Treatment: The condition applied to the subjects. When there is one factor, the treatments and the levels are the same.

Control. Control refers to steps taken to reduce the effects of other variables (i.e., variables other than the independent variable and the dependent variable). These variables are called **lurking variables**.

Control involves making the experiment as similar as possible for subjects in each treatment condition. Three control strategies are control groups, placebos, and blinding.

Control group. A control group is a group that receives no treatment

Placebo. A fake or dummy treatment.

Blinding: Not telling subjects whether they receive the placebo or the treatment

Double blinding: neither the researchers or the subjects know who gets the treatment or placebo

Randomization. Randomization refers to the practice of using chance methods (random number tables, flipping a coin, etc.) to assign subjects to treatments.

Replication. Replication refers to the practice of assigning each treatment to many experimental subjects.

Bias: when a method systematically favors one outcome over another.

Types of design:

Completely randomized design With this design, subjects are randomly assigned to treatments.

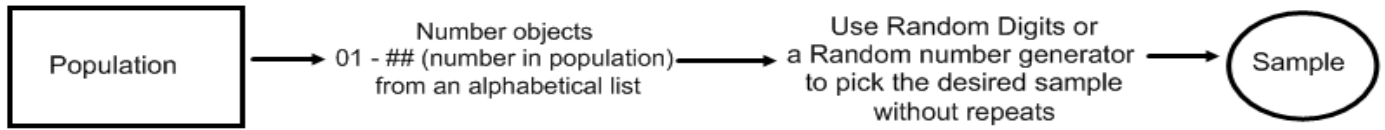
Randomized block design, the experimenter divides subjects into subgroups called **blocks**. Then, subjects within each block are randomly assigned to treatment conditions. Because this design reduces variability and potential confounding, it produces a better estimate of treatment effects.

Matched pairs design is a special case of the randomized block design. It is used when the experiment has only two treatment conditions; and subjects can be grouped into pairs, based on some blocking variable. Then, within each pair, subjects are randomly assigned to different treatments. **In some cases** you give two treatments to the same experimental unit. That unit is their own matched pair!

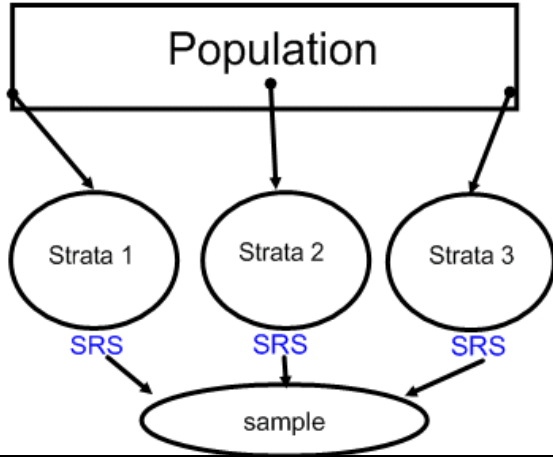
Part II in Pictures:

Sampling Methods

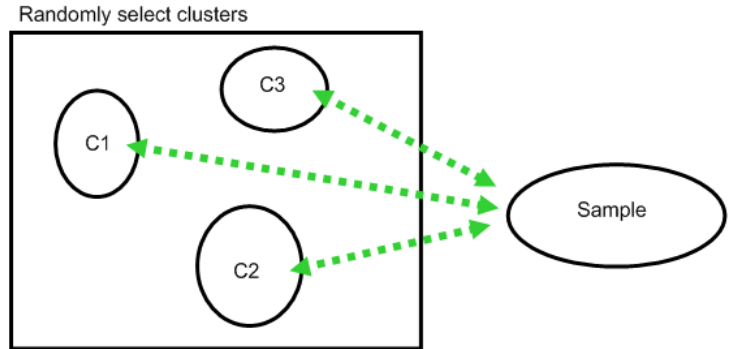
Simple Random Sample: Every group of n objects has an equal chance of being selected. (Hat Method!)



Stratified Random Sampling:
Break population into strata (groups)
then take an SRS of each group.



Cluster Sampling:
Randomly select clusters then take all
Members in the cluster as the sample.

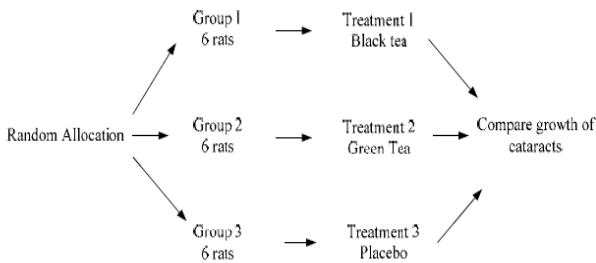


Systematic Random Sampling:
Select a sample using a system, like selecting every
third subject.

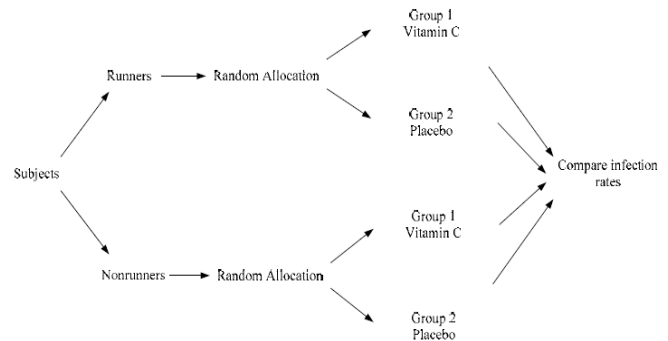


Experimental Design:

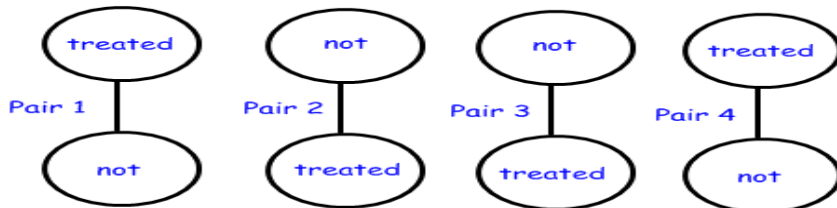
Completely Randomized Design:



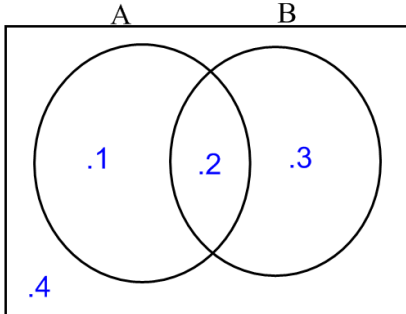
Randomized Block Design:



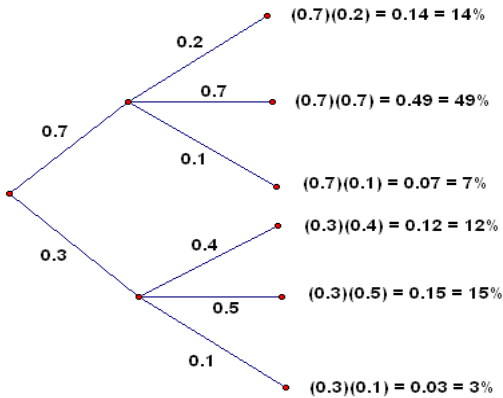
Matched Pairs Design:



Part III: Probability and Random Variables:

<p>Counting Principle: Trial 1: a ways Trial 2: b ways Trial 3: c ways ... The there are a x b x c ways to do all three. $0 \leq P(A) \leq 1$ $1 - P(A) = P(A^c)$</p>	<p>A and B are disjoint or mutually exclusive if they have no events in common. Roll two die: DISJOINT rolling a 9 rolling doubles Roll two die: not disjoint rolling a 4 rolling doubles</p>	<p>A and B are independent if the outcome of one does not affect the other. Mutually Exclusive events CANNOT BE Independent</p>	
---	---	--	---

For Conditional Probability use a TREE DIAGRAM:



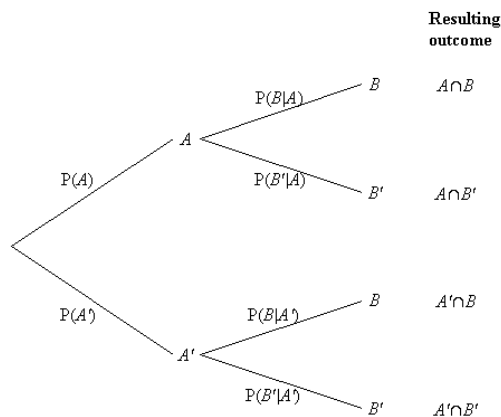
$P(A) = 0.3$
 $P(B) = 0.5$
 $P(A \cap B) = 0.2$
 $P(A \cup B) = 0.3 + 0.5 - 0.2 = 0.6$
 $P(A|B) = 0.2/0.5 = 2/5$
 $P(B|A) = 0.2 / 0.3 = 2/3$

For Binomial Probability:

- Look for x out of n trials**
1. Success or failure
 2. Fixed n
 3. Independent observations
 4. p is the same for all observations

$P(X=3)$ Exactly 3
 use binompdf(n,p,3)
 $P(X \leq 3)$ at most 3
 use binomcdf(n,p,3) (Does 3,2,1,0)
 $P(X \geq 3)$ at least 3 is $1 - P(X \leq 2)$
 use $1 - \text{binomcdf}(n,p,2)$

Normal Approximation of Binomial:
 for $np \geq 10$ and $n(1-p) \geq 10$
 the X is approx $N(np, \sqrt{np(1-p)})$



Discrete Random Variable: has a countable number of possible events (Heads or tails, each .5)
 Continuous Random Variable: Takes all values in an interval: (EX: normal curve is continuous)
 Law of large numbers. As n becomes very large $\bar{x} \rightarrow \mu$

Linear Combinations:

$\mu_{a+bx} = a + b\mu_x$

$\mu_{X+Y} = \mu_x + \mu_y$

$\sigma_{a+bx}^2 = b^2 \sigma_x^2$

$\sigma_{X+Y}^2 = \sigma_x^2 + \sigma_y^2$ $\sigma_{X-Y}^2 = \sigma_x^2 + \sigma_y^2$

Geometric Probability:

- Look for # trial until first success**
1. Success or Failure
 2. X is trials until first success
 3. Independent observations
 4. p is same for all observations

$P(X=n) = p(1-p)^{n-1}$
 μ is the expected number of trails until the first success or $\frac{1}{p}$

$\sigma^2 = \frac{1-p}{p^2}$

$P(X > n) = (1-p)^n = 1 - P(X \leq n)$

Sampling distribution: The distribution of all values of the statistic in all possible samples of the same size from the population.

Central Limit Theorem: As n becomes very large the sampling distribution for \bar{x} is approximately NORMAL

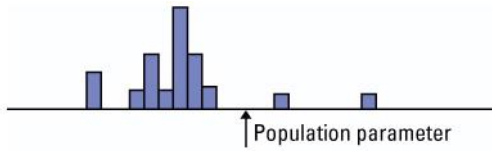
Use ($n \geq 30$) for CLT

Low Bias: Predicts the center well

High Bias: Does not predict center well

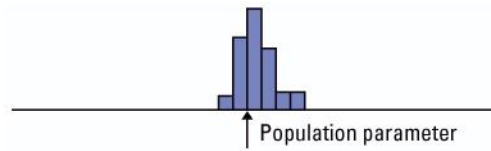
Low Variability: Not spread out

High Variability: Is very spread out



(a)

High bias, High Variability



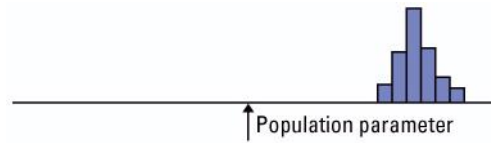
(b)

Low Bias, Low Variability



(c)

Low Bias, High Variability



(d)

High Bias, Low Variability

See other sheets for Part IV

ART is my BFF

Type I Error: Reject the null hypothesis when it is actually True

Type II Error: Fail to reject the null hypothesis when it is False.

ESTIMATE – DO A CONFIDENCE INTERVAL

EVIDENCE - DO A TEST

Paired Procedures

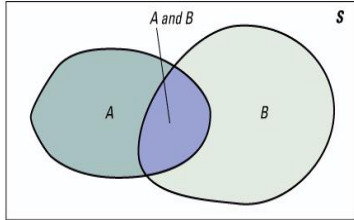
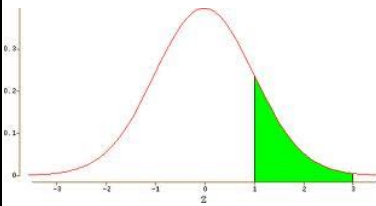
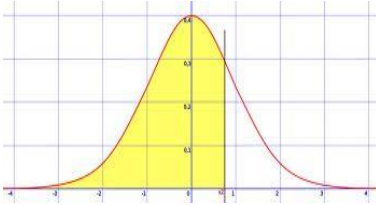
- Must be from a matched pairs design:
- Sample from one population where each subject receives two treatments, and the observations are subtracted. **OR**
- Subjects are matched in pairs because they are similar in some way, each subject receives one of two treatments and the observations are subtracted

Two Sample Procedures

- Two independent samples from two different populations **OR**
- Two groups from a randomized experiment (each group would receive a different treatment) Both groups may be from the same population in this case but will randomly receive a different treatment.

Major Concepts in Probability

For the expected value (mean, μ_x) and the σ_x or σ_x^2 of a probability distribution use the formula sheet

Binomial Probability	Simple Probability (and, or, not):
<p>Fixed Number of Trials Probability of success is the same for all trials Trials are independent</p> <p>If X is B(n,p) then (ON FORMULA SHEET) Mean $\mu_x = np$ Standard Deviation $\sigma_x = \sqrt{np(1-p)}$ For Binomial probability use $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ or use: Exactly: $P(X = x) = \text{binompdf}(n, p, x)$ At Most: $P(X \leq x) = \text{binomcdf}(n, p, x)$ At least: $P(X \geq x) = 1 - \text{binomcdf}(n, p, x-1)$ More than: $P(X > x) = 1 - \text{binomcdf}(n, p, x)$ Less Than: $P(X < x) = \text{binomcdf}(n, p, x-1)$</p> <p>You may use the normal approximation of the binomial distribution when $np \geq 10$ and $n(1-p) \geq 10$. Use then mean and standard deviation of the binomial situation to find the Z score.</p>	<p>Finding the probability of multiple simple events. Addition Rule: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ Multiplication Rule: $P(A \text{ and } B) = P(A)P(B A)$</p> <p>Mutually Exclusive events CANNOT be independent A and B are independent if the outcome of one does not affect the other. A and B are disjoint or mutually exclusive if they have no events in common. Roll two die: DISJOINT rolling a 9 rolling doubles</p> <div style="text-align: center;">  </div> <p>Roll two die: NOT disjoint rolling a 4 rolling doubles</p> <p>Independent: $P(B) = P(B A)$ Mutually Exclusive: $P(A \text{ and } B) = 0$</p>
Geometric Probability	Conditional Probability
<p>You are interested in the amount of trials it takes UNTIL you achieve a success. Probability of success is the same for each trial Trials are independent</p> <p>Use simple probability rules for Geometric Probabilities.</p> <p>$P(X=n) = p(1-p)^{n-1}$ $P(X > n) = (1-p)^n = 1 - P(X \leq n)$ μ_x is the expected number of trails until the first success or $\frac{1}{p}$</p>	<p>Finding the probability of an event given that another even has already occurred.</p> <p>Conditional Probability: $P(B A) = \frac{P(A \cap B)}{P(A)}$</p> <p>Use a two way table or a Tree Diagram for Conditional Problems. Events are Independent if $P(B A) = P(B)$</p>
Normal Probability	
<p>For a single observation from a normal population</p> <p style="text-align: center;">$P(X > x) = P(z > \frac{x - \mu}{\sigma})$ $P(X < x) = P(z < \frac{x - \mu}{\sigma})$</p> <div style="display: flex; justify-content: space-around;">   </div> <p>To find $P(x < X < y)$ Find two Z scores and subtract the probabilities (upper - lower)</p> <p>Use the table to find the probability or use normalcdf(min,max,0,1) after finding the z-score</p>	<p><u>For the mean of a random sample of size n from a population.</u></p> <p>When $n > 30$ the sampling distribution of the sample mean \bar{x} is approximately Normal with:</p> <p>$\mu_{\bar{x}} = \mu$ $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$</p> <p>If $n < 30$ then the population should be Normally distributed to begin with to use the z-distribution.</p> <p style="text-align: center;">$P(\bar{X} > x) = P(z > \frac{\bar{x} - \mu}{\sigma/\sqrt{n}})$ $P(\bar{X} < x) = P(z < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}})$</p> <p>To find $P(x < X < y)$ Find two Z scores and subtract the probabilities (upper - lower) Use the table to find the probability or use normalcdf(min,max,0,1) after finding the z-score</p>

Binomial Probability

Mr. K is shooting three point jump shots. Mr. K has a career shooting percentage of 80%. Mr. K is going to shoot 30 three pointers during a practice session.
 X: number of threes made, X is B(30, 0.6)

$$\mu_x = np = 30(.6) = 18$$

$$\sigma_x = \sqrt{np(1-p)} = \sqrt{30(.60)(.40)} = 2.683$$

The probability that Mr. K makes exactly 20 is:
 $P(X = 20) = \text{binompdf}(30, 0.6, 20) = 0.1152$

The probability that Mr. K makes at most 20 is:
 $P(X \leq 20) = \text{binomcdf}(30, 0.6, 20) = 0.8237$

The probability the Mr. K makes at least 20 is:
 $P(X \geq 20) = 1 - \text{binomcdf}(30, 0.6, 19) = 1 - 0.7085 = 0.2915$

Simple Probability (and, or, not):

$P(A) = 0.3$
 $P(B) = 0.5$
 $P(A \cap B) = 0.2$
 $P(A \cup B) = 0.3 + 0.5 - 0.2 = 0.6$
 $P(A | B) = 0.2 / 0.5 = 2/5$
 $P(B | A) = 0.2 / 0.3 = 2/3$
 $P(A^c) = 1 - 0.3 = 0.7$

Geometric Probability

The population of overweight manatees is known to be 40%
 You select a random Manatee and weigh it, and then you repeat the selection until one is overweight.

Find the probability that the fifth manatee you choose is overweight.

$$P(X = 5) = (\text{notover})^4 (\text{over}) = (0.60)^4 (0.40) = .05184$$

Find the probability that it takes more than five attempts to find an overweight manatee.

$$P(X > 5) = (\text{notoverweight})^5 = (0.60)^5 = 0.07776$$

How many manatees would you expect to choose before you found one to be overweight?

$$\mu_x = \frac{1}{p} = \frac{1}{0.4} = 2.5$$

Conditional Probability with a Tree Diagram

Of adult users of the Internet:
 29% are 18-29
 47% are 30-49
 24% are over 50

47% of the 18-29 group chat
 21% of the 30-49 group chat
 7% of the 50 and over group chat.

Find the probability that a randomly selected adult chats

Age	Chat?	Probability
A ₁	C	0.1363*
A ₁	C ^c	0.1537
A ₂	C	0.0987*
A ₂	C ^c	0.3713
A ₃	C	0.0168*
A ₃	C ^c	0.2232

Normal Probability

The weight of manatees follows a normal distribution with a mean weight of 800 pounds and a standard deviation of 120 pounds.

Find the probability that a randomly selected Manatee weighs more than 1000 pounds:
 X is N(800,120)

$$P(X > 1000) = P(z > \frac{1000-800}{120}) = P(z > 1.67) = 0.0475$$

Find the probability that a random sample of 50 manatees has a mean weight more than 1000 pounds:

$$P(\bar{X} > 1000) = P(z > \frac{1000-800}{120/\sqrt{50}}) = P(z > 11.79) \approx 0$$

Even if you did not know the population was normal you could use CLT and assume the sampling distribution is approximately normal.

Conditional Probability with a two way table:

Table 6.1 Grades awarded at a university, by school

	Grade Level			Total
	A	B	Below B	
Liberal Arts	2,142	1,890	2,268	6,300
Engineering and Physical Sciences	368	432	800	1,600
Health and Human Services	882	630	588	2,100
Total	3,392	2,952	3,656	10,000

$P(\text{A grade} | \text{liberal arts course}) = 2142 / 6300$
 $P(\text{Liberal arts course} | \text{A Grade}) = 2142 / 3392$
 $P(\text{B Grade} | \text{Engineering and PS}) = 432 / 1600$
 $P(\text{Engineering and PS} | \text{B Grade}) = 432 / 2952$

Mutually Exclusive vs. Independence

You just heard that Dan and Annie who have been a couple for three years broke up.

This presents a problem, because you're having a big party at your house this Friday night and you have invited them both. Now you're afraid there might be an ugly scene if they both show up.

When you see Annie, you talk to her about the issue, asking her if she remembers about your party.

She assures you she's coming. You say that Dan is invited, too, and you wait for her reaction.

If she says, "That jerk! If he shows up I'm not coming. I want nothing to do with him!", they're **mutually exclusive**.

If she says, "Whatever. Let him come, or not. He's nothing to me now.", they're **independent**.

Mutually Exclusive and Independence are two very different ideas

<p style="text-align: center;"><u>Mutually Exclusive (disjoint):</u> $P(A \text{ and } B) = 0$</p> <p>Events A and B are mutually exclusive if they have no outcomes in common. That is A and B cannot happen at the same time.</p> <p>Example of mutually exclusive (disjoint): A: roll an odd on a die B: roll an even on a die</p> <p>Odd and even share no outcomes $P(\text{odd and even}) = 0$ Therefore, they are mutually exclusive.</p> <p>Example of not mutually exclusive (joint): A: draw a king B: draw a face card</p> <p>King and face card do share outcomes. All of the kings are face cards. $P(\text{king and face card}) = 4/52$ Therefore, they are not mutually exclusive.</p>	<p style="text-align: center;"><u>Independence:</u> $P(B) = P(B A)$</p> <p>Events A and B are independent if knowing one outcome does not change the probability of the other. That is knowing A does not change the probability of B.</p> <p>Examples of independent events: A: draw an ace B: draw a spade</p> <p>$P(\text{Spade}) = 13/52 = 1/4$ $P(\text{Spade} \text{Ace}) = 1/4$ Knowing that the drawn card is an ace does not change the probability of drawing a spade</p> <p>Examples that are dependent (not independent): A: roll a number greater than 3 B: roll an even</p> <p>$P(\text{even}) = 3/6 = 1/2$ $P(\text{even} \text{greater than } 3) = 2/3$ Knowing the number is greater than three changes the probability of rolling an even number.</p>
---	---

<p style="text-align: center;">Mutually Exclusive events cannot be independent</p> <p>Mutually exclusive and dependent</p> <p>A: Roll an even B: Roll an odd</p> <p>They share no outcomes and knowing that it is odd changes the probability of it being even.</p>	<p style="text-align: center;">Independent events cannot be Mutually Exclusive</p> <p>Independent and not mutually exclusive</p> <p>A: draw a black card B: draw a king</p> <p>Knowing it is a black card does not change the probability of it being a king and they do share outcomes.</p>	<p style="text-align: center;">Dependent Events may or may not be mutually exclusive</p> <p>Dependent and mutually exclusive</p> <p>A: draw a queen B: draw a king Knowing it is a queen changes the probability of it being a king and they do not share outcomes.</p> <p>Dependent and not mutually exclusive</p> <p>A: Face Card B: King Knowing it is a face card changes the probability of it being a king and they do share outcomes.</p>
--	---	--

<p style="text-align: center;">If events are mutually exclusive then: $P(A \text{ or } B) = P(A) + P(B)$</p> <p>If events are not mutually exclusive use the general rule: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$</p>	<p style="text-align: center;">If events are independent then: $P(A \text{ and } B) = P(A)P(B)$</p> <p>If events are not independent then use the general rule: $P(A \text{ and } B) = P(A)P(B A)$</p>
---	--

Interpretations from Inference

Interpretation for a Confidence Interval:

I am C% confident that the true parameter (mean μ or proportion p) lies between # and #.

INTERPRET IN CONTEXT!!

Interpretation of C% Confident:

Using my method, If I sampled over and over again, C% of my intervals would contain the true parameter (mean μ or proportion p).

NOT: The parameter lies in my interval C% of the time. It either does or does not!!

If $p < \alpha$ I **reject** the null hypothesis H_0 and I have **sufficient/strong** evidence to support the alternative hypothesis H_a

INTERPRET IN CONTEXT in terms of the alternative.

If $p > \alpha$ I **fail to reject** the null hypothesis H_0 and I have **insufficient/poor** evidence to support the alternative hypothesis H_a

INTERPRET IN CONTEXT in terms of the alternative.

Evidence Against H_0

P-Value

"Some" $0.05 < \text{P-Value} < 0.10$

"Moderate or Good" $0.01 < \text{P-Value} < 0.05$

"Strong" $\text{P-Value} < 0.01$

Interpretation of a p-value:

The probability, assuming the null hypothesis is true, that an observed outcome would be as extreme or more extreme than what was actually observed.

Duality: Confidence intervals and significance tests.

If the hypothesized parameter lies **outside** the C% confidence interval for the parameter I can REJECT H_0

If the hypothesized parameter lies **inside** the C% confidence interval for the parameter I FAIL TO REJECT H_0

Power of test:

The probability that at a fixed level α test will reject the null hypothesis when an alternative value is true.

Confidence Intervals

<p><u>One Sample Z Interval</u> Use when estimating a single population mean and σ is known Conditions: -SRS -Normality: CLT, stated, or plots -Independence and $N \geq 10n$ Interval: $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ CALCULATOR: 7: Z-Interval</p>	<p><u>One Sample t Interval</u> Use when estimating a single mean and σ is NOT known [Also used in a <u>matched paired design</u> for the mean of the difference: <i>PAIRED t PROCEDURE</i>] Conditions: -SRS -Normality: CLT, stated, or plots -Independence and $N \geq 10n$ Interval: $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$ df = n-1 CALCULATOR: 8: T-Interval</p>	<p><u>One Proportion Z Interval</u> Use when estimating a single proportion Conditions: -SRS -Normality: $n\hat{p} \geq 10, n(1 - \hat{p}) \geq 10$ -Independence and $N \geq 10n$ Interval: $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$ CALCULATOR: A: 1-PropZInt</p>
<p><u>Two Sample Z Interval</u> σ known (RARELY USED) Use when estimating the difference between two population means and σ is known Conditions: -SRS for both populations -Normality: CLT, stated, or plots for both populations -Independence and $N \geq 10n$ for both populations Interval: $(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ CALCULATOR: 9: 2-SampZInt</p>	<p><u>Two Sample t Interval</u> σ unknown Use when estimating the difference between two population means and σ is NOT known Conditions: -SRS for both populations -Normality: CLT, stated, or plots for both populations -Independence and $N \geq 10n$ for both populations. Interval: $(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ Use lower n for df (df = n-1) or use calculator CALCULATOR: 0: 2-SampTInt</p>	<p><u>Two Proportion Z Interval</u> Use when estimating the difference between two population proportions. Conditions: -SRS for both populations -Normality: $n_1\hat{p}_1 \geq 10, n_1(1 - \hat{p}_1) \geq 10$ $n_2\hat{p}_2 \geq 10, n_2(1 - \hat{p}_2) \geq 10$ -Independence and $N \geq 10n$ for both populations. Interval: $(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$ CALCULATOR: B: 2-PropZInt</p>
<p align="center"><u>Confidence interval for Regression Slope:</u></p> <p>Use when estimating the slope of the true regression line Conditions: 1. Observations are independent 2. Linear Relationship (look at residual plot) 3. Standard deviation of y is the same(look at residual plot) 4. y varies normally (look at histogram of residuals) Interval: $b \pm t^* SE_b$ df = n - 2 CALCULATOR: LinRegTInt Use technology readout or calculator for this confidence interval.</p>		

Hypothesis Testing

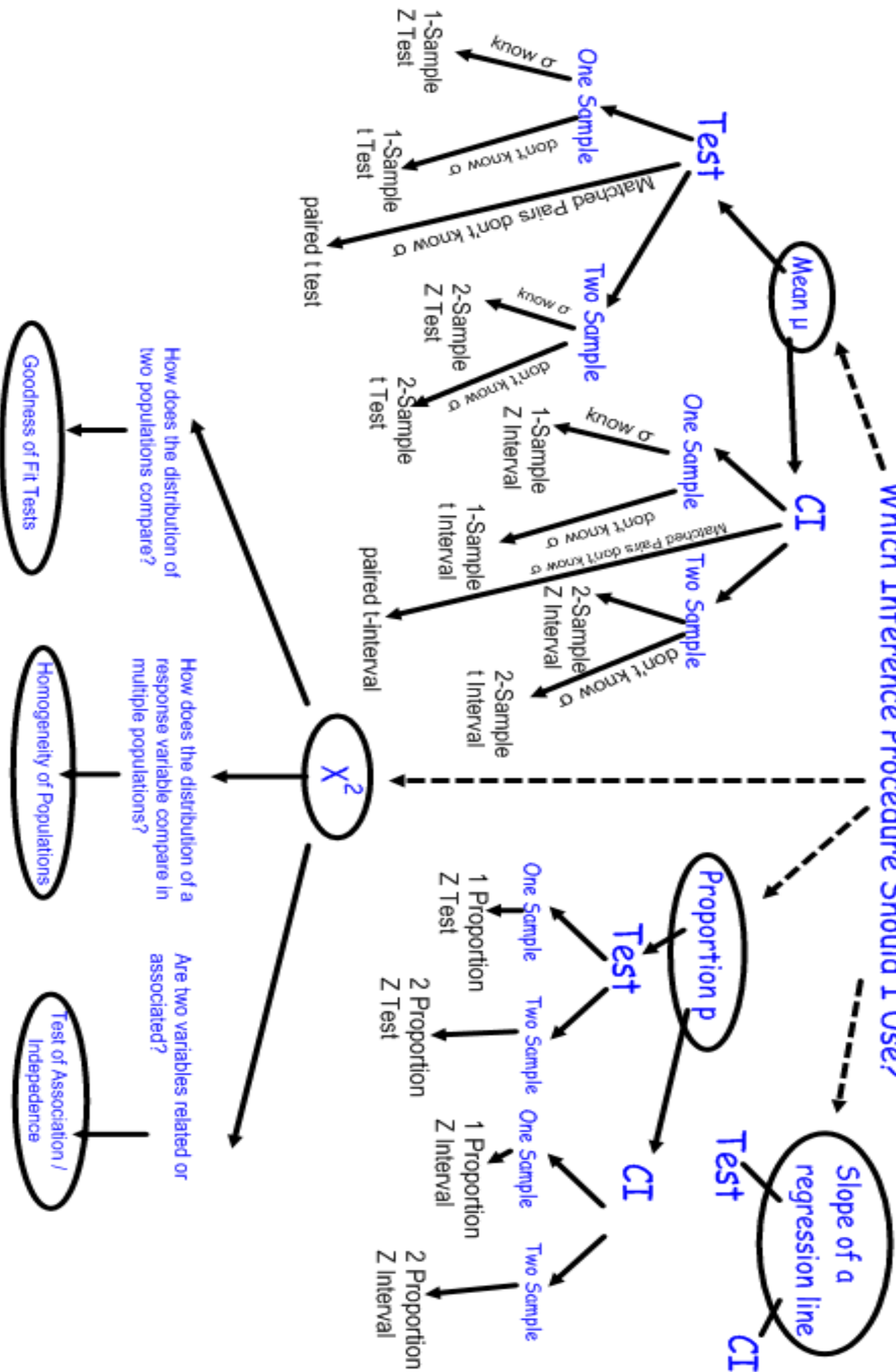
<p style="text-align: center;"><u>One Sample Z Test</u></p> <p>Use when testing a mean from a single sample and σ is known $H_0: \mu = \#$ $H_a: \mu \neq \#$ -or- $\mu < \#$ -or- $\mu > \#$ Conditions: -SRS -Normality: Stated, CLT, or plots -Independence and $N \geq 10n$ Test statistic: $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$</p> <p style="text-align: center;">CALCULATOR: 1: Z-Test</p>	<p style="text-align: center;"><u>One Sample t Test</u></p> <p>Use when testing a mean from a single sample and σ is NOT known [Also used in a <u>matched paired design</u> for the mean of the difference: <i>PAIRED t PROCEDURE</i>] $H_0: \mu = \#$ $H_a: \mu \neq \#$ -or- $\mu < \#$ -or- $\mu > \#$ Conditions: -SRS -Normality: Stated, CLT, or plots -Independence and $N \geq 10n$ Test statistic: $df = n-1$ $t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$</p> <p style="text-align: center;">CALCULATOR: 2: T-Test</p>	<p style="text-align: center;"><u>One Proportion Z test</u></p> <p>Use when testing a proportion from a single sample $H_0: p = \#$ $H_a: p \neq \#$ -or- $p < \#$ -or- $p > \#$ Conditions: -SRS -Normality: $np_0 \geq 10, n(1-p_0) \geq 10$ -Independence and $N \geq 10n$ Test statistic: $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$</p> <p style="text-align: center;">CALCULATOR: 5: 1-PropZTest</p>
<p style="text-align: center;"><u>Two Sample Z test</u> <u>σ known (RARELY USED)</u></p> <p>Use when testing two means from two samples and σ is known $H_0: \mu_1 = \mu_2$ $H_a: \mu_1 \neq \mu_2$ -or- $\mu_1 < \mu_2$ -or- $\mu_1 > \mu_2$ Conditions: -SRS for both populations -Normality: Stated, CLT, or plots for both populations -Independence and $N \geq 10n$ for both populations. Test statistic: $z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$</p> <p style="text-align: center;">CALCULATOR: 3: 2-SampZTest</p>	<p style="text-align: center;"><u>Two Sample t Test</u> <u>σ unknown</u></p> <p>Use when testing two means from two samples and σ is NOT known $H_0: \mu_1 = \mu_2$ $H_a: \mu_1 \neq \mu_2$ -or- $\mu_1 < \mu_2$ -or- $\mu_1 > \mu_2$ Conditions: -SRS for both populations -Normality: Stated, CLT, or plots for both populations -Independence and $N \geq 10n$ for both populations. Test statistic: use lower n for df ($df = n-1$) or use calculator. $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$</p> <p style="text-align: center;">CALCULATOR: 4: 2-SampTTest</p>	<p style="text-align: center;"><u>Two Proportion Z test</u></p> <p>Use when testing two proportions from two samples $H_0: p_1 = p_2$ $H_a: p_1 \neq p_2$ -or- $p_1 < p_2$ -or- $p_1 > p_2$ Conditions: -SRS for both populations -Normality: $n_1 \hat{p}_c \geq 10$ $n_1(1 - \hat{p}_c) \geq 10$ $n_2 \hat{p}_c \geq 10$ $n_2(1 - \hat{p}_c) \geq 10$ Independence and $N \geq 10n$ for both populations. Test statistic: $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c(1 - \hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$</p> <p style="text-align: center;">CALCULATOR: 6: 2-PropZTest</p>
<p style="text-align: center;"><u>χ^2 – Goodness of Fit</u> L1 and L2</p> <p>Use for categorical variables to see if one distribution is similar to another H_0: The distribution of two populations are not different H_a: The distribution of two populations are different Conditions: -Independent Random Sample -All expected counts ≥ 1 -No more than 20% of Expected Counts < 5 Test statistic: $df = n-1$ $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$</p> <p style="text-align: center;">CALCULATOR: D: χ^2GOF-Test(on 84)</p>	<p style="text-align: center;"><u>χ^2 – Homogeneity of Populations</u> r x c table</p> <p>Use for categorical variables to see if multiple distributions are similar to one another H_0: The distribution of multiple populations are not different H_a: The distribution of multiple populations are different Conditions: -Independent Random Sample -All expected counts ≥ 1 -No more than 20% of Expected Counts < 5 Test statistic: $df = (r-1)(c-1)$ $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$</p> <p style="text-align: center;">CALCULATOR: C: χ^2-Test</p>	<p style="text-align: center;"><u>χ^2 – Independence/Association</u> r x c table</p> <p>Use for categorical variables to see if two variables are related H_0: Two variables have no association (independent) H_a: Two variables have an association (dependent) Conditions: -Independent Random Sample -All expected counts ≥ 1 -No more than 20% of Expected Counts < 5 Test statistic: $df = (r-1)(c-1)$ $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$</p> <p style="text-align: center;">CALCULATOR: C: χ^2-Test</p>
<p><u>Significance Test for Regression Slope:</u> $H_0: \beta = 0$ $H_a: \beta \neq \#$ -or- $\beta < \#$ -or- $\beta > \#$ Conditions: 1. Observations are independent 2. Linear Relationship (look at residual plot) 3. Standard deviation of y is the same (look at residual plot) 4. y varies normally (look at histogram of residuals) CALCULATOR: LinRegTTest Use technology readout or the calculator for this significance test $t = \frac{b}{SE_b}$ $df = n-2$</p>		

Notation and Interpretations

IQR	Inner Quartile Range
\bar{x}	Mean of a sample
μ	Mean of a population
s	Standard deviation of a sample
σ	Standard deviation of a population
\hat{p}	Sample proportion
p	Population proportion
s^2	Variance of a sample
σ^2	Variance of a population
M	Median
Σ	Summation
Q_1	First Quartile
Q_3	Third Quartile
Z	Standardized value – z test statistic
z^*	Critical value for the standard normal distribution
t	Test statistic for a t test
t^*	Critical value for the t-distribution
$N(\mu, \sigma)$	Notation for the normal distribution with mean and standard deviation
r	Correlation coefficient – strength of linear relationship
r^2	Coefficient of determination – measure of fit of the model to the data
$\hat{y} = a + bx$	Equation for the Least Squares Regression Line
a	y-intercept of the LSRL
b	Slope of the LSRL
(\bar{x}, \bar{y})	Point the LSRL passes through
$y = ax^b$	Power model
$y = ab^x$	Exponential model
SRS	Simple Random Sample
S	Sample Space
P(A)	The probability of event A
A^c	A complement
P(B A)	Probability of B given A
\cap	Intersection (And)
U	Union (Or)
X	Random Variable
μ_x	Mean of a random variable
σ_x	Standard deviation of a random variable
σ_x^2	Variance of a random variable
B(n,p)	Binomial Distribution with observations and probability of success
$\binom{n}{k}$	Combination n taking k
pdf	Probability distribution function
cdf	Cumulative distribution function
n	Sample size
N	Population size
CLT	Central Limit Theorem
$\mu_{\bar{x}}$	Mean of a sampling distribution
$\sigma_{\bar{x}}$	Standard deviation of a sampling distribution
df	Degrees of freedom
SE	Standard error
H_0	Null hypothesis-statement of no change
H_a	Alternative hypothesis- statement of change
p-value	Probability (assuming H_0 is true) of observing a result as large or larger than that observed
α	Significance level of a test. P(Type I) or the y-intercept of the true LSRL
β	P(Type II) or the true slope of the LSRL
χ^2	Chi-square test statistic

z-score (z)	The number of standard deviations an observation is above/below the mean
slope (b)	The change in predicted y for every unit increase on x
y-intercept (a)	Predicted y when x is zero
r (correlation coefficient)	strength of linear relationship. (Strong/moderate/weak) (Positive/Negative) linear relationship between y and x.
r^2 (coefficient of determination)	percent of variation in y explained by the LSRL of y on x.
variance (σ^2 or s^2)	average squared deviation from the mean
standard deviation (σ or s)	measure of variation of the data points from the mean
Confidence Interval (#,#)	I am C% confident that the true parameter (mean μ or proportion p) lies between # and #.
C % Confidence (Confidence level)	Using my method, If I sampled repeatedly, C% of my intervals would contain the true parameter (mean μ or proportion p).
$p < \alpha$	Since $p < \alpha$ I reject the null hypothesis H_0 and I have sufficient/strong evidence to conclude the alternative hypothesis H_a
$p > \alpha$	Since $p > \alpha$ I fail to reject the null hypothesis H_0 and I have do not have sufficient evidence to support the alternative hypothesis H_a
p-value	The probability, assuming the null hypothesis is true, that an observed outcome would be as or more extreme than what was actually observed.
Duality-Outside Interval Two sided test	If the hypothesized parameter lies outside the $(1 - \alpha)\%$ confidence interval for the parameter I can REJECT H_0 for a two sided test.
Duality-Inside Interval Two sided test	If the hypothesized parameter lies inside the $(1 - \alpha)\%$ confidence interval for the parameter I FAIL TO REJECT H_0 for a two sided test.
Power of the test	The probability that a fixed level test will reject the null hypothesis when an alternative value is true
standard error (SE) in general	Estimates the variability in the sampling distribution of the sample statistic.
standard deviation of the residuals (s from regression)	A typical amount of variability of the vertical distances from the observed points to the LSRL
standard error of the slope of the LSRL (SE_b)	This is the standard deviation of the estimated slope. This value estimates the variability in the sampling distribution of the estimated slope.

Which Inference Procedure Should I Use?



CONFIDENCE EXAMPLE

A researcher believes that treating seeds with certain additives before planting can enhance the growth of plants. An experiment to investigate this is conducted in a greenhouse. From a large number of Roma tomato seeds, 24 seeds are randomly chosen and 2 are assigned to each of 12 containers. One of the 2 seeds is randomly selected and treated with the additive. The other seed serves as a control. Both seeds are then planted in the same container. The growth, in centimeters, of each of the 24 plants is measured after 30 days. These data were used to generate the partial computer output shown below. Graphical displays indicate that the assumption of normality is not unreasonable.

	N	Mean	StDev	SE Mean
Control	12	15.989	1.098	0.317
Treatment	12	18.004	1.175	0.339
Difference	12	-2.015	1.163	0.336

- (a) Construct a confidence interval for the mean difference in growth, in centimeters, of the plants from the untreated and treated seeds. Be sure to interpret this interval.
- (b) Based only on the confidence interval in part (a), is there sufficient evidence to conclude that there is a significant mean difference in growth of the plants from untreated seeds and the plants from treated seeds? Justify your conclusion.

STEPS TO ANSWER CONFIDENCE INTERVAL QUESTION CORRECTLY

Solution

Part (a):

Step 1: Identify appropriate confidence interval by name or by formula.

One sample confidence interval for a mean (of the differences)

$$OR \quad \bar{x}_d \pm t_{n-1}^* \frac{s_d}{\sqrt{n}}$$

Step 2: Check appropriate conditions.

Assume the population of differences in growth is normally distributed. The information provided in the stem of the problem suggests that this condition is met. Because the 24 seeds were randomly chosen and randomly assigned to the containers, the differences are independent.

Step 3: Correct mechanics.

The 95% confidence interval for the mean difference in growth is

$$-2.015 \pm 2.201 \frac{1.163}{\sqrt{12}} = -2.015 \pm (2.201)(0.336) = -2.015 \pm 0.7389$$

or (-2.7539, -1.2761).

Step 4: Interpret the confidence interval in context.

We are 95% confident that the mean difference in the growth of the untreated and treated seeds is between -2.7539 and -1.2761.

Part (b):

Step 1: Identify a correct pair of hypotheses.

$H_0 : \mu_d = 0$ versus $H_a : \mu_d \neq 0$, where μ_d is the mean difference in the untreated and treated seeds.

Step 2: State the correct conclusion in context.

Since the 95% confidence interval does not include zero, the null hypothesis can be rejected at the $\alpha = 0.05$ significance level. In other words, we have statistically significant evidence at the $\alpha = 0.05$ level that there is a mean difference in the growth of untreated and treated seeds.

Hypothesis Testing Example

6. Regulations require that product labels on containers of food that are available for sale to the public accurately state the amount of food in those containers. Specifically, if milk containers are labeled to have 128 fluid ounces and the mean number of fluid ounces of milk in the containers is at least 128, the milk processor is considered to be in compliance with the regulations. The filling machines can be set to the labeled amount. Variability in the filling process causes the actual contents of milk containers to be normally distributed. A random sample of 12 containers of milk was drawn from the milk processing line in a plant, and the amount of milk in each container was recorded.

- (a) The sample mean and standard deviation of this sample of 12 containers of milk were 127.2 ounces and 2.1 ounces, respectively. Is there sufficient evidence to conclude that the packaging plant is not in compliance with the regulations? Provide statistical justification for your answer.

STEPS TO ANSWER HYPOTHESIS TESTING QUESTION CORRECTLY

Solution

Part (a):

Step 1: State a correct pair of hypotheses.

$$H_0 : \mu = 128 \text{ fluid ounces versus } H_a : \mu < 128 \text{ fluid ounces}$$

Step 2: Identify a correct test (by name or by formula) and checks appropriate conditions.

One sample t -test for a mean

$$\text{OR } t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Condition: The random sample is taken from a normal population. (This information is stated in the stem so it does not need to be repeated here.)

Step 3: Use correct mechanics, including the value of the test statistic, degrees of freedom, and p -value (or rejection region)

$$\text{Test Statistic: } t = \frac{127.2 - 128}{2.1 / \sqrt{12}} = \frac{-0.8}{0.6062} = -1.3192$$

$$p\text{-value: } P(T_{11d.f.} < -1.3192) = 0.1070$$

Step 4: Using the result of the statistical test, state a correct conclusion in the context of the problem.

Since the p -value = 0.1070 is greater than any reasonable significance level, say $\alpha = .05$, we do not have statistically significant evidence to refute the claim that the company is in compliance with the regulations. That is, we cannot reject the null hypothesis that the mean quantity of milk in 12 containers is at least 128 fluid ounces.

If both an α and a p -value are given, the linkage is implied. If no α is given, the solution must be explicit about the linkage by giving a correct interpretation of the p -value or explaining how the conclusion follows from the p -value.

If the p -value in step 3 is incorrect but the conclusion is consistent with the computed p -value, step 4 can be considered as correct.

Keyboarding Tip Sheet for the 2020 AP Statistics Exam

Students may complete their responses for the 2020 AP Statistics Exam either by uploading a photo of their handwritten response or by typing on a computer or other electronic device. This Keyboarding Guide provides standard ways of entering common expressions using a standard keyboard.

Tip #1: Use the characters available on your keyboard and avoid special characters.

Examples of how some common notation can be keyboarded:

- μ : mu
- \bar{x} : x-bar
- \hat{p} : p-hat
- σ : sigma
- α : alpha
- β : beta
- Q_3 : Q_3 **OR** Q3 **OR** Q sub 3
- μ_x : mu_X **OR** mu sub X
- \hat{p}_1 : p-hat_1 **OR** p-hat sub 1
- H_0 : H_0 **OR** null hypothesis
- H_a : H_a **OR** alternative hypothesis

Examples of how some relationships and operations can be keyboarded:

- \pm : +/-
- \leq : <=
- \geq : >=
- \neq : not equal
- Binomial coefficient:
 - $\binom{8}{3}$: C(8,3) **OR** 8 choose 3
- Exponent:
 - $\left(\frac{1}{2}\right)^{12}$: (1/2)^12
 - r^2 : r^2 **OR** r-squared
- Root:
 - $\frac{5}{\sqrt{30}}$: 5/sqrt(30)

Tip #2: As always, be careful with parentheses to communicate your intended order of operations. You may need to use parentheses or brackets more frequently than when writing by hand.

The equation

$$z = \frac{2.7 - 3.2}{\frac{0.9}{\sqrt{42}}}$$

could be keyboarded as

$$z = (2.7 - 3.2) / (0.9 / \text{sqrt}(42))$$

In the expression

$$(0.45)^{3-1} (0.55)$$

the position of the exponent serves to group the expression in the exponent, in this case, "3-1". When keyboarding, you must show parentheses around the expression in the exponent, such as

$$(0.45)^{(3-1)} * (0.55)$$

Tip #3: Avoid abbreviations and shorthand to ensure intended understanding.

Some examples include:

- **Do Not indicate intervals as "x-y".** The 95 percent confidence interval is 2.5-5.5 could be interpreted as the value of the difference 2.5-5.5 instead of "from 2.5 to 5.5". Say, "The 95 percent confidence interval is from 2.5 to 5.5."
- **Do Not use abbreviations that are not standard.** For example, the abbreviation "SRS" should not be used to represent "stratified random sampling" because it is widely used to represent "simple random sampling." Spell out all terms other than standard abbreviations.
- **Do Not use notation that might imply a different operation or notation than you intend.** For example, don't write binomial coefficients in a way that might look like division. $\binom{8}{3}$ should not be written 8/3. Use "C(8,3)" or "8 choose 3" instead.
- **Do Not use calculator language unless all inputs are labeled.** For example, `normalcdf(lowerbound=4.5, upperbound=5.2, mu=4.1, sigma=1.5)`.



Calculator ID #:

Choose 2nd MEM,
#1 About
ID****_****_****

Been Playing Games?

Run DEFAULTS to reset calculator. 2nd MEM, #7
Reset, #2 Defaults, #2
Reset

To Plot Histograms and Box-Whisker Plots:

- Place data in Lists: STAT → EDIT
- Set up plot information: STAT PLOT #1 <ENTER>
Highlight ON, choose symbol for histogram, XList: L₁
OR choose symbol for box-whisker, Freq: 1
- Graph: ZOOM #9 - TRACE to see values on graph
- Xscl under WINDOW controls width of bars on histogram.
An integer value is easiest to read.

To Get Statistical Information:

- Place data in Lists: STAT → EDIT
- Engage 1-Variable Statistics: STAT → CALC #1 1-VAR STATS
- On Home Screen indicate list containing the data: 1-VAR STATS L₁

\bar{x} = mean

s_x = the sample standard deviation

σ_x = the population standard deviation

n = the sample size (# of pieces of data)

Q_1 = data at the first quartile

med = data at the median
(second quartile)

Q_3 = data at the third quartile

Diagnostics ON: must be ON to see correlation coefficient, r .

- MODE – StatDiagnostics: ON
- CATALOG, ALPHA D, DiagnosticOn, ENTER, ENTER

To Get Scatter Plots and Regressions

(Linear, Quadratic, Exponential, Power, etc)

- Place data in Lists: STAT → EDIT
- Graph scatter plot: STAT PLOT #1 <ENTER> Choose ON.
Choose the symbol for scatter plot, choose L₁, L₂, choose mark
- To graph, choose: ZOOM #9
- To get regression equation: STAT → CALC #4 Lin Reg($ax+b$)
(or whichever regression is needed)
- On Home Screen: LinReg($ax+b$) L₁, L₂, Y₁
- to see graph – GRAPH

To get Y₁ to appear:
VARS → Y-VARS
Choose
FUNCTION, Y₁
OR ALPHA F4

To Get Residuals: After preparing a regression equation (using L₁ and L₂), residuals are stored in a list called RESID.

To plot residuals:

- Go to top of L₃, press ENTER.
- Go to LIST (2nd STAT) – choose #7 RESID, press ENTER.
- Go to STAT PLOT, Plot 1, ON
- Type: first icon (scatter plot)
- XList: L₁ YList: L₃
- ZOOM 9:ZoomStat

Normal Distributions DISTR(2nd VARS)

- normalcdf** (lower, upper, mean, s.d.) *Finds prob. on cumulative interval.*
• to enter ∞ , use 10⁹⁹ or 1 EE 99.
- normalpdf**(x , mean, s.d.) *Graphs the normal distribution.*
• Window: Xmin = mean – 3 s.d.; Xmax = mean + 3 s.d.; Xscl = s.d.
Ymin = 0; Ymax = 1/(2 s.d.); Yscl = 0
- ShadeNorm**(lower, upper, mean, s.d.) *To see area and % under curve.*
• must graph using normalpdf first, or you won't see your shading.
- invnorm**(percentage, mean, s.d.)
• use when you know percentile and want to find the associated score.

Student-t Distributions DISTR(2nd VARS)

- tpdf** (x , df) *Probability density func. (graph only)*
• enter into Y=, x = variable, df (degrees freedom) > 0
- tcdf** (lower, upper, df) *Distribution probability*
• between lowerbound & upperbound, df > 0
- invT**(left tail area, df)
• not available on TI-83 models
(These commands are rarely, if ever, used at this level.)

Binomial Distributions DISTR(2nd VARS)

- binompdf** (#trials (n), prob. of success (p), # successes desired (r))
• used for a specific number of desired successes (> 0).
• if desired # not given, returns list of prob. 0 to # trials
- binomcdf**(# trials, prob. of success, # successes desired)
• finds prob. of up to # of successes desired
• if desired # not given, returns list of cumulative probs.

Geometric Distributions DISTR(2nd VARS)

- geometpdf** (prob. of success, specific trial #)
• finds prob. of a success on the specified trial #
 - geometcdf** (prob. of success, specific trial #)
• find prob. of success on, or before, specified trial #
- In both cases, the specified trial number can be a real number or a list of real numbers.
These can be tricky, so keep math formula handy.

Math Formula:

$$(1-p)^{r-1} \cdot p$$

$$p = \text{prob. success}$$

$$r = r^{\text{th}} \text{ trial}$$

Generating Random Numbers

Calculators and computers use a formula to generate “random numbers” which are called “pseudo-random”.

- Generate Random Integers (1 at a time):
MATH → PRB #5 randInt(
randInt (starting value, ending value)
- Generate Random Integers (several at a time):
randInt(starting value, ending value, # to be shown)
- Generate Random Integers in a List
randInt(0,10,100) → L₁
puts 100 integers between 0 and 100 inclusive in List 1
- To prevent random numbers from repeating, choose:
randIntNoRep(

- Generate rand numbers (not integers)

rand (generates random numbers between 0 and 1)

rand*12 (generates random numbers between 0 and 12)

rand(10)*12 → L₁ (generates 10 random numbers between 0 and 12 and stores them in List 1)

- Re-Seeding the Generator: To prevent the random list from always starting from the same number, you need to re-seed the rand command, such as **5** → **rand** (and then continue as you wish)

- Generate random numbers from Normal Distribution model

randNorm(mean, s.d.,) one at a time (not integers)

randNorm(mean, s.d., # to be shown) shows several at a time

Stat vs Data: • given actual data choose **Data** • given summary statistics (mean, s.d.), choose **Stats**.

Inferential Testing STAT (TESTS)

- Z-Test**
 - tests for one unknown pop. mean when pop. s.d. is known.
 - Use:* (1) pop. s.d. is known, (2) sample mean is known, (3) don't know pop. mean, (4) to test sample mean with some value
- T-Test**
 - test for one unknown pop. mean when pop. s.d. unknown
 - Use:* (1) sample mean is known, (2) don't know pop. mean, (3) to test sample mean with some value
- 2-SampleZTest**
 - test comparing 2 means when both pop. s.d. are known.
 - it is unusual to know BOTH pop. s.d.
 - Draw shows z-score and p-value
- 2-SampleTTest**
 - test comparing 2 means when both pop. s.d. are unknown.
 - Use:* (1) Both sample means and s.d. are known, (2) don't know pop. means, (3) to test sample mean with some value
- 1-PropZTest** (null hypothesis, # of successes (x), sample size (n), type of alt. hypothesis, display option)
 - computes a test for one proportion of successes
 - calculates z-score, p-value and proportion for sample pop.
 - if given p-hat instead of # of successes, x, calculate x by multiplying p-hat by n and rounding to nearest integer.
- 2-PropZTest** (# of successes both, both counts)
 - Test comparing 2 proportions of successes.
 - Use:* (1) working with 2 populations with different values of n where both proportions of success are known, (2) to test if there is a statistical difference.
- Chi-Square Test** (assesses goodness of fit between observed values and those expected)
 - requires observed and expected data in matrix form
 - X^2 -Test (matrix observed data, matrix expected data, display)
- Chi-Square GOF Test** (*goodness of fit*)
 - X^2 GOF-Test [works with lists]
 - use for simple random sampling, 1 categorical variable, and expected frequency of at least 5.

LinRegTTest STAT (TESTS)

- computes linear regression on data, and a t test on the value of slope and correlation coefficient
- residuals are created and stored in RESID
- use to test the degree of strength of the relationship

LinRegTInt

Confidence interval for linear regression slope coefficient b

- computes linear regression T confidence interval for the slope coefficient b. If the confidence interval contains 0, this is insufficient evidence that the data exhibits a linear relationship.

Chi-Square Distribution DISTR(2nd VARS)

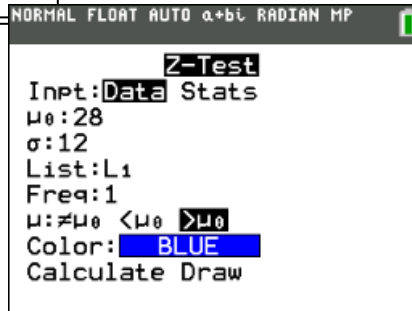
- X^2 pdf (x,df) [yields probability density function value – plots χ^2 curve with x as the variable]

The mean of a chi-square distribution equals the number of degrees of freedom of the distribution.

- X^2 cdf (lower bound, upper bound, df) computes the X^2 -distribution probability on interval [finds area under a chi-square distribution given the degrees of freedom] $P(\text{lower bound} < X^2 < \text{upper bound})$

Using Test Editors:

- Select *Data* or *Stats* input
 - select *Data* to enter data lists
 - select *Stats* to enter statistics such as mean, s.d., number
- Enter values for arguments
 - μ_0 = hypothesized value of population mean being tested
 - σ = known pop. s.d. (> 0)
 - List = name of list containing data
 - Freq = name of list containing frequency, defaults to 1
- Select *alternative hypothesis*
 - select first option for Z-test
 - select second for 2-SampTTest
 - select third for 2-PropZTest
- Select *Calculate* or *Draw* output/display option
 - Calculate* shows test calculations on the home screen Will be only choice for a Confidence Level
 - Draw* shows a graph (automatic window adjustment)



Confidence Intervals (CI) STAT (TESTS)

Calculates confidence interval for an unknown proportion of successes.

- ZInterval**
 - computes CI for unknown pop. mean with known s.d
 - assume population distribution is normal
 - be sure to highlight Calculate before hitting Enter
- TInterval**
 - computes CI for unknown pop. mean with unknown s.d
 - use when sample mean and s.d. are known
 - assume population distribution is normal
- 2-SampZInt**
 - computes CI for difference between 2 pop. means when both s.d. are known (which is quite unusual).
 - depends upon user-specified confidence level
- 2-SampTInt**
 - computes CI for difference between 2 pop. means when both s.d. are unknown.
 - use when both sample means and s.d. are known
 - assume samples are normally distributed
 - depends upon user-specified confidence level
- 1-PropZInt**
 - computes CI for unknown proportion of successes
 - use when sample size and # of successes are known
 - depends upon user-specified confidence level
- 2-PropZInt**
 - computes CI for difference between proportion of successes in 2 populations.
 - use when 2 samples have different # of successes
 - depends upon user-specified confidence level

ANOVA STAT (TESTS)

One-way analysis of variance.

ANOVA(L1, L2, L3, L4)

- computes a one-way analysis of variance for comparing the means of two to 20 populations (compares means).
- determines an F ratio to show if the means are significantly different from one list to another
- SS = sum of squares
- MS = mean squares