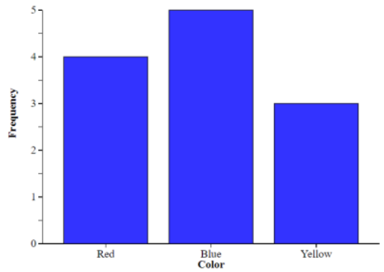
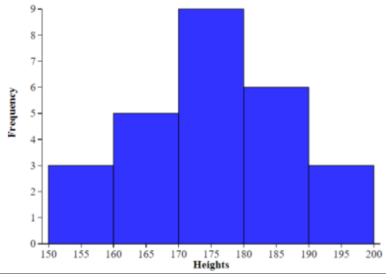
Unit 1 – Exploring One Variable Data

Bar Graphs
*Categorical



Histograms
*Quantitative, Discrete or Continuous



Stem & Leaf
*Quantitative
*Discrete

```
1 | 8 9
2 | 1 1 5 6 7
3 | 2 5 5 8 9
4 | 3 3 4
```

Data

KEY: 4|4 = 44

Split Stem & Leaf
*Quantitative
*Discrete

```
1 | 8 9
2 | 1 1
2 | 5 6 7
3 | 2
3 | 5 5 8 9
4 | 3 3 4
```

Data

KEY: 4|4 = 44

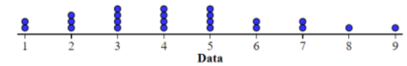
Back to Back
*Quantitative, Discrete

```
  Males     Females
      1 | 1 | 0 2
  5 4 3 | 2 | 5 6 7
6 5 2 2 | 3 | 1 2 7
    3 1 | 4 | 3
```

Data

KEY: 4|3 = 43

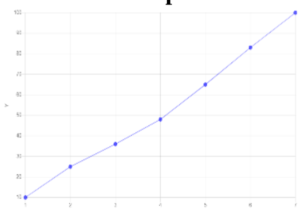
Dotplot *Discrete



Describing Distributions
S: Shape
O: Outliers
C: Center
S: Spread

Compare Distributions
Remember SOCS and use comparative words like less than, greater than, or similar to

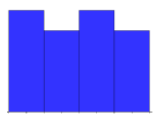
Cumulative Graph



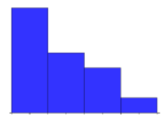
x-axis is data

y-axis is cumulative relative frequency (percentiles)

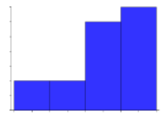
Shape



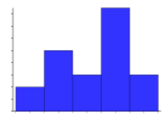
Uniform



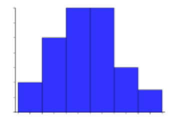
Right Skewed



Left Skewed

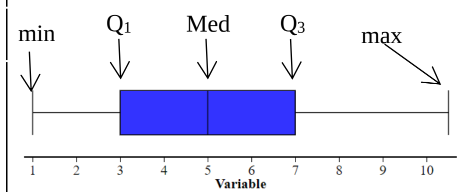


Bimodal



Unimodal Symmetric

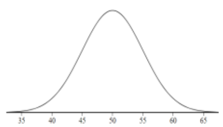
5 number summary **Boxplots**



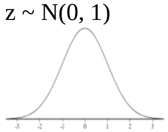
Emperical Rule:
68 – 95 – 99.7

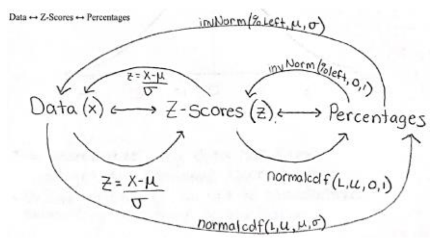
$$z = \frac{x - \mu}{\sigma}$$

x ~ N(50, 5)



Standard Normal
z ~ N(0, 1)





Center
Mean: $\bar{x}$
Median: Med

Spread
Range: max – min
IQR: $Q_3 - Q_1$
Std dev: S_x

Nonresistant:
$\bar{x}$, S_x, Range

Resistant:
Med, IQR

Outliers:
$Q_3 + 1.5(IQR) = UB$
$Q_1 - 1.5(IQR) = LB$
$\bar{x} \pm 2 \cdot S_x$

Outside these boundaries is an outlier

A modified boxplot will show outliers as dots if they go past the boundaries in the outlier test.

Percentile: percent below

Linear Transformation: ax + b
+b will change centers
ax will change centers and spreads

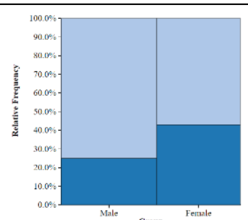
Unit 2 – Exploring Two Variable Data

	Male	Female	Total
Agree	10	15	*25*
Disagree	30	20	*50*
Total	*40*	*35*	*75*

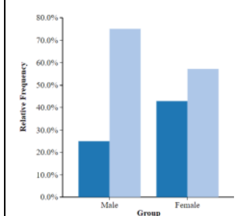
Conditional Proportions
- Agree | Male = 10/40
- Disagree | Male = 30/40

Marginal Proportions
- % that agree = 25/75
- % of females = 35/75

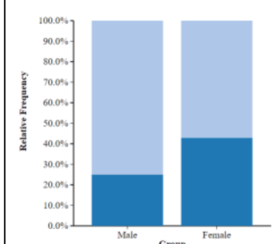
Mosaic Plot
x axis is distributed based on gender totals



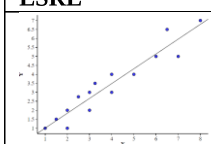
Side by Side



Segmented



LSRL



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

$$\text{slope} = b = r\left(\frac{s_y}{s_x}\right)$$

$$\text{y-int} = a = \bar{y} - b\bar{x}$$

Always passes through $(\bar{x}, \bar{y})$

Correlation: r → how close do the points fall to a linear pattern?
$-1 \le r \le 1$
direction of slope
switching x & y or changing units does not affect r

Coefficient of Determination: r^2 → % of variation in y [context] can be explained by the LSRL of y on x [context]

Describe Scatterplot: Direction, Form, and Strength
There is a strong/moderate, positive/negative, linear relationship between x [context] and y [context]

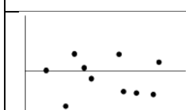
Slope: For every 1 unit increase in x, we predict an increase/decrease of [slope] in y.

	Coeff	SE	T	P
Predictor	-20	4.5	-6	0.000
x-value	6.5	1.3	5	0.000

s = 8 R-sq = 93.2%

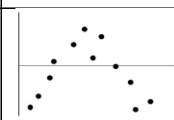
Equation:
$\hat{y} = -20 + 6.5x$
Correlation: $r = \sqrt{.932}$

Good Residual Plot



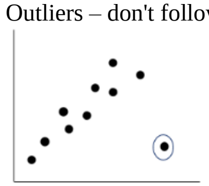
Scattered; no pattern

Bad Residual Plot

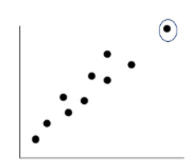


linear model not appropriate

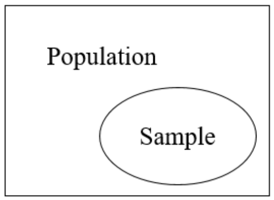
Outliers – don't follow pattern



Influential Points – influence LSRL



Unit 3 – Collecting Data



Population Parameters: p, μ, σ
Sample Statistics: , $\hat{p}$, $\bar{x}$, S_x

Observational Study: treatments are not imposed & generalizations about the population can be made

Experiment: treatments are imposed & cause & effect can be inferred

SRS: Every population member has an equal chance at being selected

Stratified: Population is sampled proportionally based on strata

Cluster: Choose an SRS of a single cluster in the population

Systematic: Every nth person is chosen

Bias: Certain responses are favored

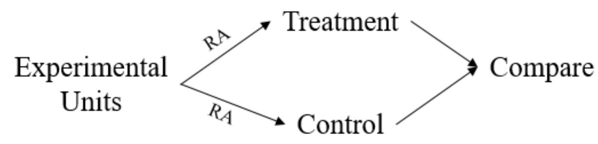
voluntary response bias, undercoverage, nonresponse, question wording bias, convenience sample

Well-Designed Experiment has:
1. Comparison
2. Random Assignment
3. Replication
4. Control

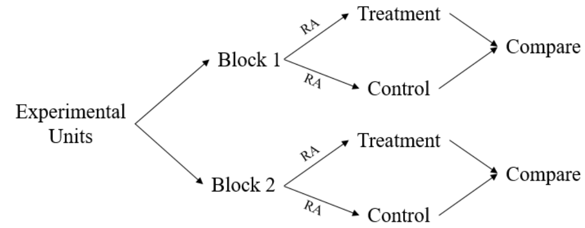
treatment, placebo, placebo-effect, double-blind, experimental units

matched pairs: each EU receives both treatments (RA: flip a coin)

Completely Randomized Design



Randomized Block Design



Ex. How to randomly assign (RA):

Place all the EU individually on a slip of paper, put all the pieces of paper in a half, shuffle, and pull out n. Those will belong to your first treatment group. The rest will go in your second treatment group.

Unit 4 – Probability

Probability: in a long series of repetitions, the proportion of times an event is successful

$$P(E) = \frac{outcomes}{sample\ space}$$

$0 \leq P(E) \leq 1$

$P(E^C) = 1 - P(E)$

Conditional Probability: Probability of A given B has already occurred

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

	Male	Female	Total
Agree	10	15	*25*
Disagree	30	20	*50*
Total	*40*	*35*	*75*

P(Male | Agree) = $\frac{10}{25}$

P(Disagree | Female) = $\frac{20}{35}$

outcome, event, simulation, relative frequency, law of large numbers, sample space, mutually exclusive

$P(A \cap B) = 0$ if mutually exclusive
$P(A \cap B) = P(A) \cdot P(B)$ if independent
$P(A \cap B) = P(A|B) \cdot P(B)$
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Test for Independence:
$P(A \cap B) = P(A) \cdot P(B)$
$P(B|A) = P(B)$

$\mu_x = E(X) = \sum x_i \cdot p_i = 1(0.2) + 3(0.1) + 5(0.3) + 6(0.4) = 4.4$

In the long run, we would expect a value of 4.4. [context of problem]

$\sigma_x = \sqrt{\sum (x_i - \mu_x)^2 \cdot p_i} = \sqrt{(1 - 4.4)^2 \cdot 0.2 + \cdots}$

Discrete RV: takes on a countable number of values

X	1	3	5	6
P(X)	0.2	0.1	0.3	0.4

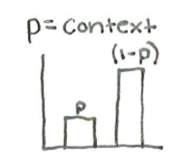
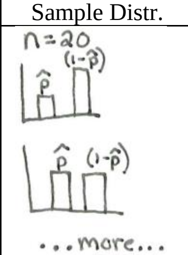
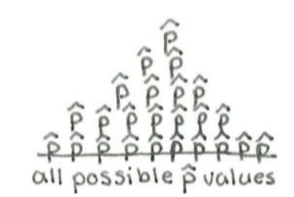
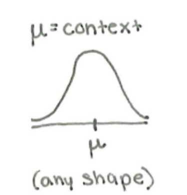
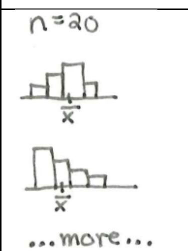
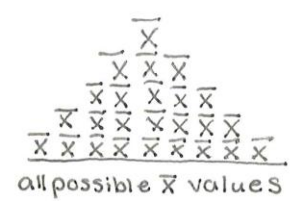
Binomial BINS	$\binom{n}{x} p^x (1-p)^{n-x}$	binompdf(n, p, x)	$\mu_x = np$	$\sigma_x = \sqrt{np(1-p)}$	successes in n trials
Geometric BITS	$(1-p)^{n-1} \cdot p$	geometpdf(p, x)	$\mu_x = \frac{1}{p}$	$\sigma_x = \frac{\sqrt{1-p}}{p}$	trials until 1st success

X ~ B(n, p)
$P(x \leq 4)$ = binomcdf(n, p, 4)
$P(x < 4)$ = binomcdf(n, p, 3)
$P(x \geq 4) = 1 - P(x \leq 3)$
$P(x > 4) = 1 - P(x \leq 4)$

Combine RV:
$E(X + Y) = E(X) + E(Y)$
$E(X - Y) = E(X) - E(Y)$
$\sigma_{X+Y} = \sigma_{X-Y} = \sqrt{\sigma_X{}^2 + \sigma_Y{}^2}$
ALWAYS ADD VARIANCES

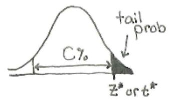
Unit 5 – Sampling Distributions

Sampling Distribution: the distribution of all possible sample statistics (either $\hat{p}$ or $\bar{x}$) of a certain size n.

	Population Distr.	Sample Distr.	Sampling Distr.	Center	Std. Dev.	Shape	Question	Answer
Proportions				If n is an SRS: $\mu_{\hat{p}} = p$	If sampling, $n \leq .1 \cdot N$ $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$	Approx. Normal if $np \geq 10$ $n(1-p) \geq 10$	p = population parameter. What is the probability that your $\hat{p}$ from a sample of n is </> a value x?	$P(\hat{p}$___x) Check Conditions $\hat{p} \sim N(\mu_{\hat{p}}, \sigma_{\hat{p}})$ $z = \frac{x - \mu_{\hat{p}}}{\sigma_{\hat{p}}} =$ ______ Probability = normalcdf(L, U, 0, 1) Interpret in context
Means				If n is an SRS: $\mu_{\bar{x}} = \mu$	If sampling, $n \leq .1 \cdot N$ $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	$n \geq 30$ Approx. Normal due to CLT $n < 30$ Sample Distr. or Population Distr. must be approx. Normal	μ = population parameter. What is the probability that your $\bar{x}$ from a sample of n is </> a value x?	$P(\bar{x}$___x) Check Conditions $\bar{x} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}})$ $z = \frac{x - \mu_{\bar{x}}}{\sigma_{\bar{x}}} =$ ______ Probability = normalcdf(L, U, 0, 1) Interpret in context

Unit 6 and 7 – Inference for Proportions and Means

Confidence Level: If you were to repeat this many times, C% of the resulting CI will contain the true population parameters
z* = invNorm(tail prob, 0, 1)
t* = invT(tail prob, df)
df = n – 1



To find n given a set ME:

$ME = z^*\sqrt{\frac{p(1-p)}{n}}$ or $ME = z^*\left(\frac{\sigma}{\sqrt{n}}\right)$

- Use past studies for p and σ estimates (or use p = 0.5)
- Always round up!

p-value = the probability that you get a statistic like this or more extreme (in the direction of H_a) if H_0 is true

		Truth	
		H_0	H_A
Decision	H_A	I	Power
	H_0	Boring	II

P(Type I) = α = Reject H_0 incorrectly
Power = Rejecting H_0 correctly
P(Type II) = β = Fail to Reject H_0 incorrectly
Power = 1 – β

Confidence Intervals

Set-Up	Conditions	Formula	Conclusion
1-sample z interval for population proportions (capture p) p = [context]	1. Random (SRS/Assign) 2. n ≤ .1N 3. $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$	$\hat{p} \pm z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ TI84: 1-PropZInt	We are ____% confidence that the interval from _____ to _____ captures the true population proportion of [context].
1-sample t interval for population means (capture μ) μ = [context]	1. Random (SRS/Assign) 2. n ≤ .1N 3. n ≥ 30 CLT n < 30 graph sample	$\bar{x} \pm t^*\left(\frac{S_x}{\sqrt{n}}\right)$ TI84: tInterval	We are ____% confidence that the interval from _____ to _____ captures the true population mean of [context].
2-sample z interval for population proportions (capture $p_1 - p_2$) p_1 = [context] p_2 = [context]	1. Both Random (SRS/Assign) 2. Both n ≤ .1N 3. Both $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$	$\hat{p}_1 - \hat{p}_2 \pm z^*\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ TI84: 2-PropZInt	We are ____% confidence that the interval from _____ to _____ captures the true difference ($p_1 - p_2$) in [context].
2-sample t interval for population means (capture $\mu_1 - \mu_2$) μ_1 = [context] μ_2 = [context]	1. Both Random (SRS/Assign) 2. Both n ≤ .1N 3. Each n ≥ 30 CLT n < 30 graph sample	$\bar{x}_1 - \bar{x}_2 \pm t^*\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ TI84: 2-SampTInt	We are ____% confidence that the interval from _____ to _____ captures the true difference ($\mu_1 - \mu_2$) in [context].

Significance Tests

Test and Hypotheses	Conditions	Formula	p-value	Conclusion p < ∝ Reject H_0 p > ∝ Fail to Reject H_0
1-sample z test for population proportions H_0: $p = p_0$ H_A: p _____ p_0 p = [context]	1. Random (SRS/Assign) 2. n ≤ .1N 3. $np_0 \geq 10$ and $n(1-p_0) \geq 10$	$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ TI84: 1-PropZTest	z = ______ normalcdf(L, U, 0, 1) p = ______	Because our p-value of ______ is </> ∝ = ____, we reject H_0/fail to reject H_0. There is/is not convincing evidence that [H_A in context]
1-sample t test for population means H_0: $\mu = \mu_0$ H_A: μ _____ μ_0 μ = [context]	1. Random (SRS/Assign) 2. n ≤ .1N 3. n ≥ 30 CLT n < 30 graph sample	$t = \frac{\bar{x} - \mu_0}{S_x/\sqrt{n}}$ TI84: tTest	t = ______ df = n – 1 tcdf(L, U, df) p = ______	Because our p-value of ______ is </> ∝ = ____, we reject H_0/fail to reject H_0. There is/is not convincing evidence that [H_A in context]
Matched Pairs **1-sample t test** for a mean difference H_0: $\mu_d = 0$ H_A: μ_d _____ 0 μ_d = [context]	1. Random (SRS/Assign) 2. n ≤ .1N 3. n ≥ 30 CLT n < 30 graph sample	$t = \frac{\bar{x}_d}{S_x/\sqrt{n}}$ TI84: tTest	t = ______ df = n – 1 tcdf(L, U, df) p = ______	Because our p-value of ______ is </> ∝ = ____, we reject H_0/fail to reject H_0. There is/is not convincing evidence that [H_A in context]
2-sample z test for population proportions H_0: $p_1 = p_2$ H_A: p_1 _____ p_2 p_1 = [context] p_2 = [context]	1. Random (SRS/Assign) 2. n ≤ .1N 3. $n_1p_C \geq 10$ $n_1(1-p_C) \geq 10$ $n_2p_C \geq 10$ $n_2(1-p_C) \geq 10$	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ $p_c = \frac{x_1 + x_2}{n_1 + n_2}$ TI84: 1-PropZTest	z = ______ normalcdf(L, U, 0, 1) p = ______	Because our p-value of ______ is </> ∝ = ____, we reject H_0/fail to reject H_0. There is/is not convincing evidence that [H_A in context]
2-sample t test for population means H_0: $\mu_1 = \mu_2$ H_A: μ_1 _____ μ_2 μ_1 = [context] μ_2 = [context]	1. Both Random (SRS/Assign) 2. Both n ≤ .1N 3. Each n ≥ 30 CLT n < 30 graph sample	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ TI84: 2-SampTTest	t = ______ df = n – 1 tcdf(L, U, df) p = ______	Because our p-value of ______ is </> ∝ = ____, we reject H_0/fail to reject H_0. There is/is not convincing evidence that [H_A in context]

Unit 8 – Inference for Categorical Data: Chi-Square

χ^2 Goodness of Fit Test	χ^2 Test for Homogeneity	χ^2 Test for Association/Independence
H_0: The claimed distribution is correct. H_A: At least one of the claimed proportions is incorrect. IN CONTEXT	**2 separate samples from 2 unique populations** H_0:There is no difference in the distribution of [context] H_A: There is a difference in the distribution of [context]	**1 sample from a single population** H_0:There is no association between ____ & ____ (They are independent) H_A: There is an association between ____ & _____ (They are not independent)
1. Random (SRS/Assign) 2. n ≤ .1N (Sampling w/o replacement) 3. All expected counts ≥ 5 (Show table of expected counts)	1. Random (SRS/Assign) 2. n ≤ .1N (Sampling w/o replacement) 3. All expected counts ≥ 5 (Show table of expected counts) expected counts in each cell = $\frac{(row\ total)(column\ total)}{table\ total}$	1. Random (SRS/Assign) 2. n ≤ .1N (Sampling w/o replacement) 3. All expected counts ≥ 5 (Show table of expected counts) expected counts in each cell = $\frac{(row\ total)(column\ total)}{table\ total}$
$\chi^2 = \sum \frac{(O-E)^2}{E}$ TI84: χ^2GOFTest χ^2 = _______ p-value = χ^2cdf(χ^2, 1e99, df) df = number of categories – 1	$\chi^2 = \sum \frac{(O-E)^2}{E}$ TI84: χ^2Test χ^2 = _______ p-value = χ^2cdf(χ^2, 1e99, df) df = (number of rows – 1) (number of columns – 1)	$\chi^2 = \sum \frac{(O-E)^2}{E}$ TI84: χ^2Test χ^2 = _______ p-value = χ^2cdf(χ^2, 1e99, df) df = (number of rows – 1) (number of columns – 1)
Because our p-value of ______ is </> ∝ = ____, we reject H_0/fail to reject H_0. There is/is not convincing evidence that [H_A in context]	Because our p-value of ______ is </> ∝ = ____, we reject H_0/fail to reject H_0. There is/is not convincing evidence that [H_A in context]	Because our p-value of ______ is </> ∝ = ____, we reject H_0/fail to reject H_0. There is/is not convincing evidence that [H_A in context]

Unit 9 – Inference for Slopes

population regression equation:	sample regression equation:
$\mu_y = \alpha + \beta x$	$\hat{y} = a + bx$

Sampling Distribution for Slope

Shape: Approx. Normal as long conditions below are met Mean: $\mu_b = \beta$ Std. Dev: $\sigma_b = \frac{\sigma}{\sigma_x/\sqrt{n}}$	σ = standard deviation of residuals (s on computer output) σ_x = standard deviation of x-values n = sample size

t-interval for β	Used to capture β, the true population slope t* = invT(tail prob, df) df = n – 2	$b \pm t^* SE_b$ $SE_b = \frac{s}{s_x\sqrt{n-1}}$ TI84: LinRegTInt	We are ____% confident that the interval from ____ to ____ captures the true population slope of the regression line. [in context] If this interval contains 0, there is no evidence of an association.
t-test for β	H_0: β = 0 H_A: β ___ 0 β = [context]	$t = \frac{b}{SE_b}$ = ______ p-value = tcdf(L, U, df) df = n – 2 TI84: LinRegTTest	Because our p-value of ______ is </> ∝ = ____, we reject H_0/fail to reject H_0. There is/is not convincing evidence that [H_A in context]

Conditions

L: Linear	**I: Independent**	**N: Normal**	**E: Equal SD**	**R: Random**
check scatterplot for a linear relationship between x & y	if sampling w/o replacement, check n ≤ .1N	Histogram of residuals should be approx.. Normal	Residual plot should have random scatter	SRS or random assignment

	Coeff	SE	T	P
Predictor	-20	4.5	-6	0.000
x-value	6.5	1.3	3.5	0.003

s = 8 R-sq = 93.2%

Equation: $\hat{y} = -20 + 6.5x$
SE_b = 1.3
t = 3.5
p-value = 0.003

H_0: β = 0
H_A: β ≠ 0
Divide by 2 to get one-sided p-value