* Confidence Interval For $p_1 - p_2$

Suppose we have two groups where we want to compare the proportion of a trait.

Let n_1 be the sample size for a sample from population 1, and y_1 = # of people with trait out of the n_1.

Let n_2 be the sample size for a sample from population 2, and y_2 = # of people with trait out of the n_2

Let $\hat{p}_1 = y_1/n_1$ = sample proportion for group 1
Let $\hat{p}_2 = y_2/n_1$ = sample proportion for group 2.

Let $\tilde{p}_1 = (y_1+1)/(n_1+2)$ = Wilson adjusted proportion for group 1
Let $\tilde{p}_2 = (y_2+1)/(n_2+2)$ = Wilson adjusted proportion for group 2

Notice the correction is split between the two groups. The idea is still the same - a CI based on $\tilde{p}$'s is strictly better than one that depends on $\hat{p}$'s.

* The $(1-\alpha)100\%$ Wilson-Adjusted CI for $p_1 - p_2$ is:

$$(\tilde{p}_1 - \tilde{p}_2) \pm Z_{\alpha/2}\sqrt{(\tilde{p}_1(1-\tilde{p}_1))/(n_1+2) + (\tilde{p}_2(1-\tilde{p}_2))/(n_2+2)}$$

Notice this is a CI for a difference in proportions, and while the bounds may be calculated as values larger than 1 or less than -1, in practice we would report any value over 1 as 1, and any value less than -1 as -1.

Similar to a CI for $\mu_1 - \mu_2$, if
i) Both bounds are positive, there is evidence to suggest $p_1 > p_2$.

ii) Both bounds are negative, there is evidence to suggest $p_2 > p_1$.

iii) The bounds contain 0, there is evidence to suggest there is no sig. difference between p_1, p_2.

Ex: For patients who suffer from migraines, two treatments were considered:
A = new drug, B = placebo, and if their pain was significantly reduced (R) or not.

	A	B
R	41	15
R^c	8	11
	$n_1 = 49$	$n_2 = 26$

a) Estimate the proportion of patients who had a significant reduction in pain, by group and overall.

$\hat{P}_r\{R|A\} = 41/49 = 0.8367$ $\hat{P}_r\{R|B\} = 15/26 = 0.576$
$\hat{P}_r\{R\} = (41+15)/(49+26) = 0.7466$

b) Find a 95% CI for the difference in proportion of pain reduction for A vs. B.

$\tilde{p}_1 = (41+1)/(49+2) = 0.824$, $\tilde{p}_2 = (15+1)/(26+2) = 0.571$
$Z_{\alpha/2} = t_{\alpha/2}$ at $d.f = \infty = t_{.05/2} = 1.96$

$(0.824 - 0.571) \pm 1.96\sqrt{(.824)(1-.824)/(49+2) + (.571)(1-.571)/(26+2)}$
$\Rightarrow (0.042, 0.464)$

c) Interpret your interval in c) in terms of the problem.

We are 95% confident that the new drug has

a higher true probability of reducing pain than the placebo by between 4.2% and 46.4%.

d) What is the largest difference in proportions you would expect with 95% confidence?
46.4% or 0.464, since that is the largest difference suggested by the CI in b).

Assumptions for a CI for $p_1 - p_2$

1) Random samples were taken from both groups
2) Groups are independent
3) y_1, y_2, $(n_1 - y_1)$, $(n_2 - y_2)$ are all ≥ 5.

Statistical vs. Practical Significance.

We have experienced that when sample size increases, error decreases, and test statistics grow larger.

This means that when we have large sample sizes, we are more likely to pick up extremely small differences in our data compared to Ho, and we are more likely to reject Ho.

What can happen is that the differences we pick up are so small that they are in a practical application, NOT significant, but statistically, they are significant.

Linear Relationships

In two categorical random variables, we asked "how does one variable effect the other?"
If we have two numeric variables, we can ask the same question, but now we start by asking about linear relationships.

Say we have random variables X and Y.

Definition: X and Y have a positive linear relationship if when X increases, Y tends to increase.

Definition: X and Y have a negative linear relationship if when X increases Y tends to decrease.

In addition, we have specific names for X and Y:

X is often the variable which we believe helps explain the behavior of Y, or the explantory variable.
Y is often the variable which we believe responds to the change in X, or the response variable.

Ex: Identify X and Y.

a) Blood Pressure and salt intake.
We believe increases in salt intake may partially explain an increase in BP, so X = salt intake, Y = BP.

b) # of Parasites and Fruit yield
We believe increases in # parasites may partially explain an increase in fruit yield, so X = # of parasites Y = fruit yield.

Correlation

Correlation between two variables means they have a linear relationship. The next question would be, how strong of a linear relationship is it?

Correlation Coefficient

Let ρ (greek letter rho) be the population correlation coefficient, and let r = sample correlation coefficient.

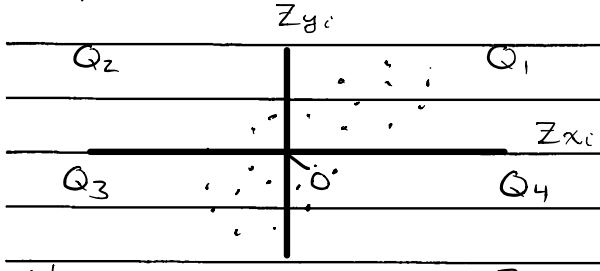
We assume the data comes in pairs, and we measure (x_i, y_i) from each subject.

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\begin{matrix}\text{z-score} \\ \text{for } x_i\end{matrix}\right)\left(\begin{matrix}\text{z-score} \\ \text{for } y_i\end{matrix}\right)$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (z_{x_i})(z_{y_i})$$

Recall a z-score standardizes data, so that it has mean 0, std. dev. 1.

Say we have a positive linear relationship. If we plot the z-scores, they will look like:



Notice when z_{x_i}, z_{y_i} are in Q_1, $(z_{x_i})(z_{y_i}) > 0$
when z_{x_i}, z_{y_i} are in Q_3, $(z_{x_i})(z_{y_i}) > 0$
when z_{x_i}, z_{y_i} are in Q_2, $(z_{x_i})(z_{y_i}) < 0$
when z_{x_i}, z_{y_i} are in Q_4, $(z_{x_i})(z_{y_i}) < 0$

But, when there is a positive linear relationship, there are a larger proportion of points in Q_1, Q_3 than Q_2, Q_4. Thus, $r = \frac{1}{n-1} \Sigma (z_{x_i}) z_{y_i} > 0$.

Similarly for a negative linear relationship.

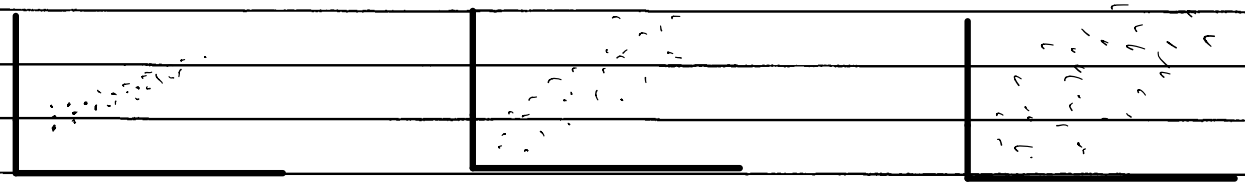
Properties about ρ or r

1) $-1 \leq r \leq 1$
2) r is unitless, and can be used to compare different datasets.
3) If $r = -1$, the relationship between X and Y is a perfect negative linear relationship, and all the points lie on a perfect line with a negative slope.
4) If $r = +1$, the relationship between X and Y is a perfect positive linear relationship, and all the points lie on a perfect line with a positive slope.
5) If $r = 0$, there is no linear relationship.

Thus, the closer r is to ± 1, the stronger the linear relationship, and the closer r is to 0, the weaker the linear relationship.

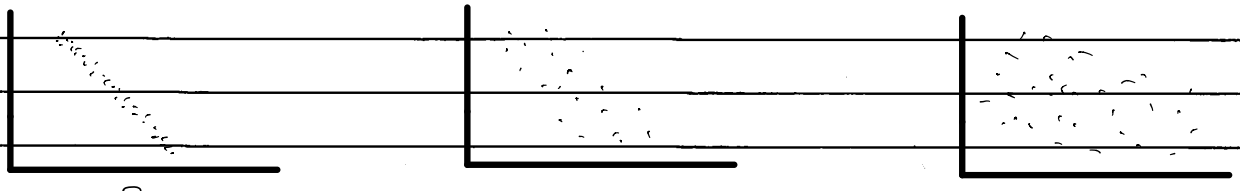
Ex:

Positive Linear



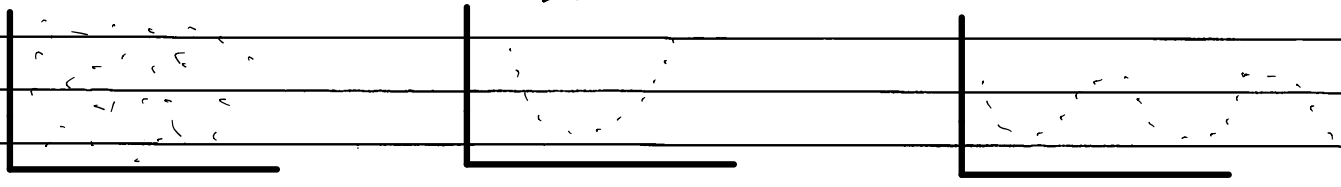
$r = .8$ $r = .4$ $r = .2$

Negative Linear



$r = -.9$ $r = -.5$ $r = -.1$

No Linear



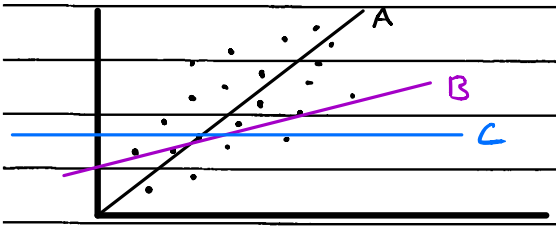
$r = 0$ $r = 0$ $r = 0$

Note: You will not have to calculate r by hand, it would be given.

✳ Linear Regression

If we believe Y tends to change when X changes, the next step would be to quantify how Y responds to a change in X.

Idea: Fit a line to our data. But which line?
Let a general line be
$\hat{y}_i = b_0 + b_1 x_i$ for pairs (y_i, x_i).
The goal is to model Y's behavior with X.



A, B, and C are all possible lines. A appears to match the data the best, but how can we tell?

Notice the "error" each line makes can be measured.
Let e_i = vertical error from y_i to the line $(b_0 + b_1 x_i)$
$= (y_i - (b_0 + b_1 x_i)) = (y_i - \hat{y}_i)$
Now, consider
overall error = $\Sigma(e_i^2)$ = Sum of Squared Errors = SSE.

The "best fit" line, or "least squares" line or "estimated regression" line is the line that minimizes SSE.

Using calculus, we can find that that line has the following slope and intercept:

$b_1 = (r)(S_y/S_x) = (\text{measure of strength of linear relationship})\left(\frac{\text{deviation in } y}{\text{deviation in } x}\right)$
which is essentially the statistical version of the algebraic Formula
Slope = (rise/run) = (change in y)/(change in x)

and then

$b_0 = \bar{y} - b_1\bar{x}$

* Using the estimated regression line

Now that we have a "best" line, we may:

1) Interpret the slope. "The estimated change in Y when X changes by one unit is b_1 on average."

2) Interpret the intercept: "The estimated average of Y when $X = 0$ is b_0".

Note: If X cannot be 0 in practice, this may not have a practical meaning.

3) Predict a value of Y at a specific $X = x^*$. We would just plug in x^* to the line:

y^* = estimated $y = b_0 + b_1 x^*$

4) Calculate the errors based on our data

$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$

Ex: A researcher believed that temperature (°F) may effect the # of cricket chirps in 1 sec. A sample from 15 different days showed

	Chirps/sec	Temp °F
mean	16.537	79.34
std dev	1.708	7.020

and $r = 0.832$

a) Identify X and Y.
Since we believe temp effects cricket chirps,
X = temp, Y = chirps

b) Find the estimated regression line
$b_1 = r(s_y/s_x) = .832(1.708/7.020) = 0.202$
$b_0 = 16.537 - 0.202(79.34) = 0.5103$

$\hat{y} = 0.5103 - 0.202x$ (or $\hat{y}_i = 0.5103 + 0.202x_i$)

c) Interpret the slope in terms of the problem.
When temperature increases by 1°F, we estimate # of chirps/sec will increase by 0.202 on average.

d) Interpret the intercept in terms of the problem, if appropriate.
When temp = 0°F, we would expect all crickets to be dead, so there is no practical interpretation.

e) If at temp 79°F we heard 14 chirps/se, what is the error based on our line? Did we over, or under estimate the value?
$e_i = 14 - (.5103 + 0.202x) = 14 - (0.5103 + 0.202(79))$
$= 14 - 16.4683 = -2.4683 \Rightarrow$ over estimated, since $16.4683 > 14$.

"True" Regression Model

We assume a "true" slope and intercept exist, but also acknowledge that the data will vary from this line. Thus, the "true model" we assume is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Y_i = i^{th} value of the r.v. Y

X_i = i^{th} value of the r.v. X

β_0 = "true" intercept based on the population

β_1 = "true" slope based on the population

ε_i = individual error for the i^{th} subject. This allows all individuals in the population to vary from the "true" line.

b_0 estimates β_0, b_1 estimates β_1, e_i estimates ε_i

Assumptions of Linear Regression

1) Y and X have a linear relationship.
2) A random sample of pairs was taken.
3) All pairs of data (x_i, y_i) are independent (or all ε_i are indep.)
4) The variance of the errors ε_i is constant.
5) The average of the errors is zero.
6) The errors are normally distributed

Note: 4, 5, 6 can be summarized as $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$

Fact: If $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$, $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2_\varepsilon)$

I.e., at any particular X value, Y is normally distributed with mean $(\beta_0 + \beta_1 X_i)$, variance σ^2_ε.

Confidence Interval for β_1

Fact: Since if $\varepsilon \sim N(0, \sigma^2_\varepsilon)$, Then $Y \sim N(\ )$, b_1 is also distributed normal.

$b_1 \sim N(\beta_1, \sigma_{b_1}^2 = \sigma_\varepsilon^2 / \Sigma(x_i - \bar{x})^2)$

Note: $\Sigma(x_i - \bar{x})^2 = S_x^2(n-1)$

But, we have to estimate σ_ε^2 with S_e^2, so in practice b_1 is distributed t with d.f. $= n-2$.

Since we know the distribution of b_1 (if our assumptions hold), we can make a CI:

A $(1-\alpha)100\%$ CI for β_1 is:

$$b_1 \pm t_{\alpha/2} \, S_e / \sqrt{S_x^2(n-1)} \quad \text{with d.f.} = n-2$$

Some general rules follow:

i) If both bounds are > 0, this suggests a significant positive linear relationship.

ii) If both bounds are < 0, this suggests a significant negative linear relationship.

iii) If the bounds contain 0, this suggests $\beta_1 = 0$ is a plausible value, and thus would suggest no significant linear relationship.

We can estimate σ_ε^2, and it has a practical meaning.

Let $SSE = SS(\text{resid}) = \sum_{i=1}^{n} e_i^2 =$ Sum of Squared Errors (or Residuals).

Let S_e estimate σ_ε^2. Then,

$S_e = \sqrt{SSE/(n-2)}$

S_e is interpreted as "a typical value of an error when estimating Y using the linear relationship with X"

Note: SSE has units of (units of Y)^2, S_e has units of Y.

Coefficient of Determination

We would also like a way to assess how well our model fits. One way to do this is by using

r^2 = coefficient of determination

$$= [\Sigma(y_i - \bar{y})^2 - \Sigma(e_i^2)] / \Sigma(y_i - \bar{y})^2 = (r)^2$$

Note:

$\Sigma(y_i - \bar{y})^2$ can be viewed as overall error if we used $\bar{y}$ to predict every y_i.

Σe_i^2 is the overall error if we used $\hat{y}_i = b_0 + b_1 X_i$ to predict y_i

Thus, $(\Sigma(y_i - \bar{y})^2 - \Sigma e_i^2)$ is the reduction in error when using regression to predict y_i, instead of using $\bar{y}$.

Finally, $[(\Sigma(y_i - \bar{y})^2 - \Sigma e_i^2)] / \Sigma(y_i - \bar{y})^2$

is interpreted as:

"The percentage of reduction in error when using the linear regression of Y and X to predict Y, instead of using the sample mean"