Data/R

Data Visualization

`geom_hist and geom_density`	distribution of numerical columns
`geom_bar`	number of occurences in a categorical col
`geom_boxplot`	shape & distribution of numerical vars
`geom_scatter + geom_line*`	numerical vs. numerical
`geom_bar`	bar plot for count of categorical vars
`geom_hline(yintercept)`	horizontal line
`geom_vline(xintercept)`	vertical line
`geom_abline(slope, intercept)`	linear function, requires
`geom_segment`	straight line between (x, y) and (xend, yend)
`geom_smooth`	plots a line/curve of best fit

*`geom_line` only makes sense with an ordering (e.g. the x-axis is year and observations connect together)

Data Manipulation

`arrange(asc(col))`	arranges *col* by ascending order
`arrange(desc(col))`	arranges *col* by descending order
`relocate(data, col, .before, .after)`	relocates a column relative to its neighbors*
`arrange(desc(col))`	arranges *col* by descending order
`slice(data, pos)`	indexes rows
`bind_rows(df1, df2, ...)`	dfs w/ same columns, concats rows
`bind_cols(df1, df2, ...)`	dfs w/ same # rows, concats cols, renames repeated cols
`semi_join(x, y, by)`	returns rows from x w/ matching val for `by` in y
`anti_join(x, y, by)`	returns rows from x w/o a match in y
`full_join(x, y, by)`	standard outer join
`left_join(x, y, by)`	standard left join, x is the left df
`right_join(x, y, by)`	standard right join, y is the right df

*specifying no neighbors moves *col* to leftmost col, specifyfing both is error

Suppose we have the following table `fish_encounters`

fish	station	seen
4842	Release	1
4842	I80_1	1
4842	Lisbon	1
4842	Rstr	1
4842	Base_TD	1
4842	BCE	1
4842	BCW	1
4842	BCE2	1
4842	BCW2	1
4842	MAE	1
4845	BCE	0

pivot_wider(fish_encounters, names_from = station, values_from = seen, values_fill = 0)

Fish	Release	I80_1	Lisbon	Rstr	Base_TD	BCE	BCW	BCE2	BCW2	MAE
1	4842	1	1	1	1	1	1	1	1	1
2	4843	1	1	1	1	1	1	1	1	1
3	4844	1	1	1	1	1	1	1	1	1
4	4845	1	1	1	1	0	0	0	0	0

Suppose we have the following table `billboard`

artist	track	date.entered	wk1	wk2	wk3	wk4	wk5	wk6	wk7
2 Pac	Baby...	2000-02-26	87	82	72	77	87	94	99
2Ge+her	The ...	2000-09-02	91	87	92	NA	NA	NA	NA
3 Doors D...	Kryp...	2000-04-08	81	70	68	67	66	57	54
3 Doors D...	Loser	2000-10-21	76	76	72	69	67	65	55
504 Boyz	Wobb...	2000-04-15	57	34	25	17	17	31	36

pivot_longer(billboard, cols = starts_with("wk"), names_to = "week", names_prefix = "wk", values_to = "rank", values_drop_na = TRUE)

artist	track	date.entered	week	rank
2 Pac	Baby Don't Cry (Keep...	2000-02-26	1	87
2 Pac	Baby Don't Cry (Keep...	2000-02-26	2	82
2 Pac	Baby Don't Cry (Keep...	2000-02-26	3	72
2 Pac	Baby Don't Cry (Keep...	2000-02-26	4	77
2 Pac	Baby Don't Cry (Keep...	2000-02-26	5	87
2 Pac	Baby Don't Cry (Keep...	2000-02-26	6	94
2 Pac	Baby Don't Cry (Keep...	2000-02-26	7	99
2Ge+her	The Hardest Part Of ...	2000-09-02	1	91
2Ge+her	The Hardest Part Of ...	2000-09-02	2	87
2Ge+her	The Hardest Part Of ...	2000-09-02	3	92

Dates & Strings

`ymd(), dmy(), ...`	converts string to datetime according to order of y-m-d
`wdate(date)`	gets the day of the week for a given date
`strc(str1, str2, ...)`	concatenates strings/vectors of strings

`str_detect(str, pattern)` TRUE if $\exists$ a substring of `str` that matches `pattern`

`str_extract(str, pat, group)` finds 1st match in `str` for `pat`, `group` takes matched pattern, returns text matching `group`

`str_extract_all(string, pattern)` returns all matches to `pattern`

`str_sub(string, start, end)` indexes into `string`

`str_count(string, pattern)` count # of matches to `pattern` in `string`

str_replace(string, pattern, replacement), str_replace_all(string, pattern, replacement) - these exist

putting color, fill, alpha, etc. outside of aes(), i.e. typically inside of geom_x() functions will set it as a constant for the whole graph

putting color, fill, alpha, etc. inside of aes() typically implies you have a column in your df (like year) that sets the groups appropriately

every geom_x() function inherits the aes() from ggplot, unless they have their own aes() which overrides the ggplot

R always prints dates as YYYY-MM-DD

Regex

`\d`	digits
`\s`	whitespace
`\w`	alphabetic and numeral
`^`	matches the start of each line
`$`	matches the end of each line
`?`	0 or 1
`+`	1 or more
`*`	0 or more
`{n}`	exactly n
`{n, }`	n or more
`{n, m}`	between n and m

Capitalizing any of the above is the complement

You can also create your own character classes using []:

`[abc]`	matches a, b, or c
`[a-z]`	matches every character between a and z
`[^abc]`	matches anything except a, b, or c
`[\^\-]`	matches ^or -

Parenthesis make groups which can be backreferenced

pattern <- "(..)\\1" #(..) is some pair of anything, and 1 takes that same pair

fruit %>% str_subset(pattern)

"banana" "coconut" "cucumber" "jujube" "papaya" "salal berry"

Basic Probability

Probability Theory

For some random variable X, $E(X) = \sum_{x=0}^{n} x * P(X = x)$.
The expected value is just the sum of each outcome multiplied by its porbability.

$Var(X) = E((X - \mu)^2)$, $\mu = E(X)$

Again, this is just multiplying the squared difference of the mean from each observation with each observation's respective probability, $sum((x - mu)^2 * p)$.

Suppose that the distribution of X is proportional with the function $g(x) = 6 - |x - 5|$.
Say that we have outcomes $1, 2, \cdots, 10$, this means $P(X = x) = a(6 - |x - 5|)$.
We know that the total number of outcomes and number of current outcomes must be proportional to the function.
The way to make the number of outcomes proportional is to find $\sum_{i=1}^{10} 6 - |i - 5|$.

To keep the possible values proportional, each probability is $\frac{j}{\sum_{i=1}^{10} 6-|i-5|}$, $j \in g(1), g(2), \cdots, g(10)$.

Binomial Distributions

Properties of Binomials

`b`	binary outcomes
`i`	independence*
`n`	fixed sample size
`s`	same probability

Sampling w/o replacement violates this (when drawing from a set of outcomes you remove outcomes sampled).

Binomial Formulas

$\mu = np$ $\quad \sigma^2 = np(1 - p)$

binom prob $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$

R Binomial Functions

`rbinom(n, size, prob)` random binomial samples

`dbinom(x, size, prob)` density fcn at `x`

`qbinom(p, size, prob)` get the smallest value in the qth quantile

`pbinom(q, size, prob)` $P(X <= $ `q`$)$

`pbinom(q, size, prob, lower.tail = T)` $1 - P(X <= $ `q`$) = p(X >$ `q`$)$

Note that $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Normal Distributions

R Normal Functions*

`pnorm(q, size, prob)` $\quad P(X <$ `q`$)$

At any given x, $X \sim N(\mu, \sigma)$, $P(X = x) = 0$.
The standard normal is $X \sim N(0, 1)$.
Any normal has its z-scores as equivalent observations in the standard normal. In other words, $X \sim N(\mu, \sigma) \implies Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$.

*includes `rnorm()`, `qnorm()` which have same functionality as the binom fcns

Suppose that $\exists X \sim Binom(n, p)$, with $np(1 - p) \geq 10$
Note the conditions this tests, p can't be too close to 0 or 1 (causes skew), and n must be sufficiently large (reduces variance).
We can approximate that binomial with $X \sim N(np, \sqrt{np(1 - p)})$.
Recall that this appoximation isn't perfect, the normal has an effect of "cutting off" the binomial distribution.
Correct for this with $P(X \leq x + .5)$ wheing finding $P(X \leq x)$, $P(X \geq x - .5)$ when finding $P(X \geq x)$

As a general rule,
65% of data 1 SD from the mean
95% of data 2 SD from the mean
99% of data 3 SD from the mean

Inference

Inference on Proportions

Formulas ($\hat{P}*$ is a random estimator for point estimate $\hat{p}$):

$Var(\hat{P})$	$Var(\frac{X}{n}) = \frac{Var(X)}{n}$
$Var(X)$	$\frac{p(1-p)}{n}$
$SE(\hat{p})$	$\sqrt{Var(\hat{P})}$
CI	$\hat{p} \pm z * SE$
z	$qnorm(1 - \frac{p}{2})*$

$*\hat{P} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$, this is still random

*where p is the desired conf interval

Agresti-Coull Method (use $\tilde{p}$ in place of $\hat{p}$)

$\tilde{p}$	$\frac{x+2}{n+4}$
$SE(\tilde{p})$	$\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$
CI	$\tilde{p} \pm z * SE$
z	$qnorm(1 - \frac{p}{2})*$

*where p is the desired conf interval

In theory this is a better estimate, still when SE is too small the CI can be too narrow.
Using $\tilde{p}$ moves the estimate closer to .5.
When $\hat{p}$ is closer to 0 or 1 than p, SE tends to be underestimated, and vice versa for $\hat{p}$ closer to .5 than p.

Hypothesis testing - determine if a result we found was due to random chance

1. Have a binomial model
2. State H_0 and H_A
3. Choose test statistic
4. Find p-value and see if it's under some α*

*Conventionally, we call $p < .05$ statistically significant and $p < .01$ highly statistically significant.

Assume H_0 is true. Now find probability we observed a certain outcome.
Suppose $H_0|X \sim Binom(n, k)$ and $H_0 : k = .5$, $H_A : k \neq .5$. We observe j successes and n observations in total. Then p is $2 * pbinom(j, n, .5)$, since the probability distribution is symmetric and $\neq$ necessitates a 2-sided test.
H_0 also assumes a binomial distribution w/ chance of success being .5.
Findings are summarized in the following way: There is strong evidence (p=0.0021) , two-sided binomial test) that the chimpanzee in this experiment will make the pro-social choice more than half the time in the long run under similar experimental conditions.

Differnce in Proportions

$\bar{p}$	$\frac{x_1+x_1}{n_1+n_2}$
$SE(\hat{p_1} - \hat{p_2})$	$\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}$
CI	$\tilde{p} \pm z * SE$
z	$qnorm(1 - \frac{p}{2})*$

*where p is the desired conf interval

Agresti-Coffe Method (use $\tilde{p}$ in place of $\hat{p}$)

$\tilde{p}$	$\frac{x+1}{n+2}$
$SE(\tilde{p_1} - \tilde{p_2})$	$\sqrt{Var(\tilde{p_1}) + Var(\tilde{p_2})}$
$Var(\tilde{p_i})$	$\frac{\tilde{p_i}*(1-\tilde{p_i})}{(n_i+2))}$
CI	$\tilde{p} \pm z * SE$
z	$qnorm(1 - \frac{p}{2})*$

Hypothesis Testing Difference in Proportions - determine if result was due to random chance

1. Have 2 binomial models
2. State H_0 and H_A
3. Choose test statistic
4. Find p-value and see if it's under some α

Say that $H_0 : p_1 - p_2 = 0$, $H_A : p_1 \neq p_2$. Say that the number of success is i_1 and i_2 respectively.
Since we test $p_1 - p_2$ the differnces will be normally distributed.
Estimate the combined probability as $\bar{p} = \frac{i_1+i_2}{n_1+n_2}$.
Again, assuming H_0 is true calculate z. $z = \frac{(\hat{p_1}-\hat{p_2})-(p_1-p_2)}{SE}$. THIS $p_1 - p_2$ IS THE $p_1 - p_2$ DEFINED BY H_0.
p is the area area under the standard normal, or $2 * P(X > z)$ in this case.
This test is called the z test for differences in proportions.

Inference on Means

Formulas:

`CI`	$z * SE$
z	$qt(p, n - 1)$
$SE(\bar{x})$	$\frac{\sigma}{\sqrt{n}}$
T	$\frac{X-\mu_0}{s/\sqrt{n}}*$
s	$\sqrt{\frac{\Sigma_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$

$*\mu_0$ is the value assumed to be μ under H_0. Interpret this as the # of SE above/below the mean of the null distribution.

The p value is the area under the t distribution WRT the T statistic, with $n-1$ degrees of freedom.
WHEN FINDING THE CI OF ANY NORMAL/T DISTRIBUTIONS THE CRITICAL VALUE IS SOME QNORM/QT ACCORDING TO THE DESIRED CONFIDENCE LEVEL.

Inference on Multiple Means

Data can be paired or unpaired. Paired data is observations that are similar, and we are interested in differences between them.
For paired data:
Consider a new distribution of the **differences** in each pair of observations. Hypothesis testing, confidence intervals, etc. are exactly the same as inference on a single mean, just on the difference between means this time.
For unpaired data:
If the variance of the 2 distributions is similar, use the 2-sample:

$SE(\bar{X}-\bar{Y})$	$SE = \sqrt{\frac{\Sigma(x_i-\bar{x})^2+\Sigma(y_i-\bar{y})^2}{n_x+n_y-2}} \cdot \sqrt{\frac{1}{n_x}+\frac{1}{n_y}}$
Statistic	$t_{obs} = \frac{(\bar{X}-\bar{Y})-(\mu_{X0}-\mu_{Y0})}{SE(\bar{X}-\bar{Y})}$
Degrees of freedom	$DF = n_x + n_y - 2$
Interval	$(\bar{X}-\bar{Y}) \pm t_{crit} * SE$

Welch when variance different:

$SE(\bar{X}-\bar{Y})$	$SE = \sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}$
Statistic	$t_{obs} = \frac{(\bar{X}-\bar{Y})-(\mu_{X0}-\mu_{Y0})}{SE(\bar{X}-\bar{Y})}$
Degrees of freedom	$DF = \frac{(s_x^2/n_x+s_y^2/n_y)^2}{(s_x^2/n_x)^2/(n_x-1)+(s_y^2/n_y)^2/(n_y-1)}$
Interval	$(\bar{X}-\bar{Y}) \pm t_{crit} * SE$

Where $\mu_{X0}-\mu_{Y0}$ is the difference in means assumed under H_0.
p process is same as before, find area under t distribution according to H_a WRT the T statistic.

T Distributions

Recall the T statistic for inference on means.
$T = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$, notice that we use s, a point estimate for σ. This introduces randomness, so T is not quite normally distributed, so we use t distribution. The standard deviation is $\frac{d}{d-2}$, $d > 2$ where d is degrees of freedom. If $d \in \mathbb{Z}$, round it down. In practice, the t distribution converges to the normal as d increases. Still, it resembles a stretched normal.

Regression

Linear Regression

$y = \beta_0 + \beta_1 x_i + \epsilon_i$ ST the line minimizes the sum squared error.
Formulas:

$r = Corr(x,y)*$	$\frac{1}{n-1}\sum_{i=1}^n \frac{x_i-\bar{x}}{s_x}\frac{y_i-\bar{y}}{s_y}$
$\hat{\beta_1}$	$r * \frac{s_y}{s_x}$
$\hat{\beta_0}$	$\bar{y} - \hat{\beta_1}\bar{x}$
$sigma(lm)$	extract SD of errors

$*r$ assumes x & y linearly related, measures strength of assumed relationship.

The regression line always goes thru $(\bar{x}, \bar{y})$.
If the residual grpah is curved, $\bar{\epsilon} \neq 0$. If there's fanning out/narrowing in residual plot, $SE(\epsilon)$ is not constant.
Confidence intervals for $E(y \mid x^*)$
This is a confidence interval for the mean y given some x^*. For example, if there were infinite observations of x^* we would be p% confident they fall into a given range (you can also think of this as a confidence interval for points regression line passes thru).

$$s_{\hat{y}} = \sqrt{\left((n-2)^{-1}\sum_{i=1}^n (y_i-\hat{y}_i)^2\right)\left(n^{-1}+\frac{(x^*-\overline{x})^2}{\Sigma_{i=1}^n (x_i-\overline{x})^2}\right)}$$

Prediction intervals for $E(y^* \mid x^*)$
This is a confidence interval for some random y^*. So not only do we account for uncertainty in β_0 and β_1, we also have randomness in ϵ. Since, by nature observations have randomness baked in which is encoded by ϵ.
As such, this interval is wider than the confidence interval.
Prediction interval is an interval that 95% of observations in the dataset will be in

$$s_{\hat{y}^*} = \sqrt{\left((n-2)^{-1}\sum_{i=1}^n (y_i-\hat{y}_i)^2\right)\left(1+n^{-1}+\frac{(x^*-\overline{x})^2}{\Sigma_{i=1}^n (x_i-\overline{x})^2}\right)}.$$

Power Law

Sometimes a relationship can only be expressed as $y = C * x^{\theta}$, this is the power law.
In this case apply a log transform to the data, that way it is normal. Any models fit on this data will predict $log_l(\mathbf{y})$, however.
The regression line is trying to predict the mean for each observation. That is, each observation will have a bit of randomness, the regression line wants to capture the mean of all the hypothetical observations. So we say
$y_i \sim N(log_l(C) + \theta x_i, \sigma)$.
To hypothesis test whether the power law is appropriate we carry out the following hypothesis test (assuming $\epsilon \sim N(0,\sigma)$):
$H_0 : \theta = 1$, $H_a : \theta \neq 1$. This is called a t-test for regression slope.
Hypothesis testing:

T	$\frac{\hat{\theta}-\theta_0}{s_{\hat{\theta}}}$
$s_{\hat{\theta}}$	$\sqrt{\frac{\Sigma_{i=1}^n \epsilon_i^2/(n-2)}{\Sigma_{i=1}^n (x_i-\bar{x})^2}}$

Use $n-2$ degrees of freedom when doing inference on regression. This has the same process as inference on a single mean.

Specifics

Inference

For any type of inference dealing w/ difference, if
$H_a : u_1 \neq u_2 \equiv u_1 - u_2 < 0 \vee u_1 - u_2 > 0$.
If you find u_1 and u_2 are not the same you can say something about which is greater. If there are any transformations done to θ in the power law. For example, $y = C * x^{-\theta}$, then the confidence intervals, etc. must reflect this. In practice, this involves multiplying the given x from the linear model by -1.
For all hypothesis tests the q functions are used to find t_crit and p functions used to find $p - value$.

Linear Regression

Suppose you are given $r = .68$, $\bar{x} = 67$, $s_x = 4$, $s_y = 30$. Given $\hat{x} = 73$ what is $\hat{y}$?
To solve find the z-score of $\bar{x}$ relative to its distribution, $z = \frac{73-67}{4} = 1.5$.
Now for some reason, $\hat{y} = z * r * s_y = .68 * 1.5 * 30$.

Power Law

Suppose we have the following relationship $y = C * x^{-\theta}$. To validate the hypothesis we need $.16 \leq \theta \leq .19$, having $\theta = .25$ will invalidate the hypothesis.
First we take the natural log to get a linear relationship
$ln(y) = ln(C) + -\theta ln(x)$. A linear model is fit and we are given the ouput $summary(lm)$. Note that $n = 83$.

Coeffs	Estimate	SE	T (assuming $\mu_{X_0} = \mu_{Y_0}$)	$P(>\lvert t\rvert)$
Intercept	-0.17977	0.08046	-2.234	0.0282
x	-0.16071	0.02405	-6.681	2.76e-09

It also gives $df = 81$.
The CI for θ: $.16071 + c(-1,1) * 0.02405 * qt(.975, 81)$.
Assuming $H_0 : \theta = .25$ and $H_a : \theta \neq .25$, the test statistic:
$(.16071 - .25)/0.02405$. The $p-value$: $pt((.16071 - .25)/0.02405, 81)$.

Suppose we are trying to prove Moore's law, that is $T = C2^{\frac{Y}{2}}$, where T is the # of transistors in a computer chip, and Y is the year.
Taking logs of both sides, $log_{10}(T) = log_{10}(C) + \frac{log_{10}2}{2}Y$, given $log_{10}2/2 = .1505$. $H_0 : Y = .1505$, $H_a : Y \neq .1505$. $n = 99$.
Suppose the test statistic is $.1716$, $p = 2 * qt(.1716, 97) = .86$. Then we conclude that the observed data is consistent with Moore's law that number of transistors in computer chips doubles every two years (p=0.86 , t-test for regression slope, 97 df).
$\hat{y} = \hat{b_0} + \hat{b_1} * x$, $\hat{b_0} = \bar{y} - \hat{b_1}\bar{x}$, $\hat{b_1} = r * \frac{s_y}{s_x}$
Given $r = .68$, $s_y = 30$, $s_x = 4$, $\bar{x} = 67$, $x = 73$. Want to find $\hat{y} - \bar{y}$.
$\hat{b_1} = .68 * (30/4) = .68 * 7.5$, $\hat{b_0} = \bar{y} - .68 * 7.5 * 67$.
$\hat{y} = (\bar{y} - .68 * 7.5 * 67) + (.68 * 7.5 * 73)$
$\hat{y} - \bar{y} = .68 * 7.5 * 73 - .68 * 7.5 * 67 = 6 * .68 * 7.5$

Confidence Intervals

Single proportion: Use $N(0,1)$ for confidence intervals, use $z = qnorm(1 - ((1-p)/2))$ with it.
Diff in proportions: $N(0,1)$ for confidence intervals, use $z = qnorm(1 - ((1-p)/2))$ with it.
Single mean: $t(n-1)$, use $qt(1 - ((1-p)/2), n-1)$
Diff in means: t(Welch's approximate degrees of freedom), use $qt(1 - ((1-p)/2), Welch's - 1)$
Linear regression: $t(n-2)$, use $z = qt(1 - ((1-p)/2), n-2)$

Hypothesis Testing

Single proportions, $Binom(n, p_{null})$ for hypothesis testing, find p using $pbinom(x, n, p_{null})$
Diff proportions, $N(0,1)$ for hypothesis testing, find p using $pnorm(z)$
Single mean, $t(n-1)$, use $pt(T, n-1)$.
Diff in means: t(Welch's approximate degrees of freedom), use $pt(T, Welch's - 1)$
Linear regression: $t(n-2)$, use $z = pt(T, n-2)$