STAT 201

Types of Claims: summary, generalization, causal, prediction   Variable: characteristic of obj. that can be measured

Variables   DF: each row is unit of obs.   Contingency Table: count or freq. of combs. of vars.

numerical / categorical   Proportions: Joint: appear in a comb. of levels of vars   Association: cond. props vary as you go from one level to next
cont / disc   ordinal / Nominal   Marginal: appear in a single level of a single var
disc   Conditional: A given B   geom_bar()   ggplot(data, mapping = aes(x=..., y=...))
mag. has meaning   Density plot: smoothed histogram   position = fill → normalized   + geom_...() + labs() + annotate()
     Violin plot: density plot as boxplot: histogram   dodge → dodged   Geometries   Aes Attributes
     Skew is dir of tail   mean is sensitive to outliers — w/ name   stack → stacked   point, bar, line,   x = loc on x-axis
$IQR = Q3 - Q1$ sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ mutate(df, col = ...)   histogram, density,   y = loc on y-axis
$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$ summarise(df, sum1 = mean(...),   filter(df, criteria)   violin, dotplot, boxplot color, alpha, size, shape,
     grouped   sum2 = median(...), ...) Σ can use &, | fill
     group_by(df, [col])   select(df, col1, col2, ...)   %in%   arrange(df, desc(col))   use |>
Correlation Coeff. $r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$ Least Sq. $\hat{y} = b_0 + b_1 x$ $\hat{y} = $ fitted(model)   — top/num. omit for asc
     $b_1 = r\frac{s_y}{s_x}$ lm(y~x, data=df) $\hat{e} = $ resid(model)   MLR: lm (formula = y ~ $x_1 + x_2$, data = df)
$r = cor(x,y)$ $b_0 = \bar{y} - b_1\bar{x}$ Residuals: categorical as indicator vars
     $\hat{e}_i = y_i - \hat{y}_i$ Bernoulli: Indicator w/ prob p   ncol() gives total num cols
Prob Hist theoretical ${}_nP_k = \frac{n!}{(n-k)!}$ $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ Binom: k successes in n identical trials   factorial(n) = n!
Emp. Hist empirical   choose(n,k) = $\binom{n}{k}$
     $\binom{n}{k}p^k(1-p)^k$ e.g. summarise(df, n = ncol())
Probability Code   replicates   replicate(n, expr)   # times   HyperG: k successes in n draws, G successes,
rep(x, times, each)   values in   a vector   geom.cd() for prob hists $\frac{\binom{G}{k}\binom{N-G}{n-k}}{\binom{N}{n}}$ N-G failures
     can be   def. 1   sample(data, size, replace)   P(B|A) = $\frac{P(A\cap B)}{P(A)}$
a vec of same len   dbinom(x, size, prob) = P(X=x)   Poisson: n >> p   P(A∪B) = P(A) + P(B) − P(A∩B)
as x   pbinom(q, size, prob) = F(X=x) $P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$ CDF: $F(x) = \sum_{y\le x} f(y)$
dhyper(x, m, n, k) = P(X=x)   rbinom (n, size, prob) samples n from Binom (n,p) $\mathbb{E}[X] = \sum x f(x)$ $\mathbb{E}[cX] = c\mathbb{E}[X]$ $\mathbb{E}[c] = c$
phyper(q, m, n, k) = F(X=x)   dpois(x, lambda) $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ $\mathbb{E}[g(x)] = \sum g(x) f(x)$
rhyper(nn, m, n, k) samples size nn   ppois(q, lambda)   Dist $\mathbb{E}$ Var $Var[X] = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
     rpois (n, lambda)   Binom   np   np(1-p)
$S_n = X_1 + X_2 + ... + X_n$ $\bar{X} = \frac{S_n}{n}$ HyperG   nG/N $Var[cX] = c^2 Var[X]$, $Var[X+c] = Var[X]$
     Poisson λ 2 $Var[X+Y] = Var[X] + Var[Y]$
$\mathbb{E}[S_n] = E[\bar{X}] = E[X_i]$ Std. Err is std. dev of   CLT: Sum & avgs. of iid goes to Normal
$Var[\bar{X}] = \frac{Var[X_i]}{n}$ statistic $X \sim N(\mu, \sigma^2)$ Interval   Bernoulli: p   p(1-p)   sample(x, prob=probs)   Sample c Population
     Area (U(0,1))   [-1,1]   .68   Bias: selection (biased samp.) statistic   pop param
dunif(x, min, max)   dnorm(x, mean, sd)   [-2,2]   .95   measurement (systematic err in measuring)
punif(q, min, max)   pnorm(q, mean, sd)   [-3,3]   .997   non-response
runif(n, min, max)   rnorm(n, mean, sd)   Sampling Distn: dist of stat upon   Variation: sampling, measurement
infer for bootstrapping   Normal by CLT   repeated sampling $SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$ $\mu = \bar{x}, \sigma = s$ 95% CI: $\pm 1.96 SE$

df |> specify(response = col) |> generate(reps = n, type = (bootstrap, draw, permute)) |> calculate (stat = ("mean", "median", "proportion",
     "diff-in-means", "diff-in-props")) |>
Hyp Testing   hypothesize(null = ("independence", "point")), visualize(), shade-p-value("diff-in-means", "diff-in-props") |>

H_0: chance process   test stat; leans on H_0 α: signif. level, thresh.   obs $\hat{p}$, directional   fit()
H_A: alternative mech.   p-val: prob of obs test-stat or   for p-val   stat   "both", "less   2 no args; optionally fits
Replication - subj to treatment gp → control   more extreme   TypeI: reject H_0, actually true, $\hat{p}$ than", "more than")   lm
covariate - var collected before administering treatment   TypeII: accept H_0, H_A acc. true, $\hat{p}$ A cause B if $!A \Rightarrow !B$ counterfact
is balanced: if approx same in treatment gp & control gp.   min $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ ATE = $E[Y(1)] - E[Y(0)]$ cobalt (needs num or logical)
$SMD(x) = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{\sigma_t^2}{2} + \frac{\sigma_c^2}{2}}}$ $R^2 = \frac{SSR}{TSS} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$ optim(f, par)   bal.tab (y ~ $x_1 + x_2$, data = df, s.d.denom = "pooled", binary = "std")
     bad   good $[0,1]$ $= \frac{RSS}{TSS} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$ z starting pt. |> plot() → love plot
Overfitting w/ Train & Test Set   Broom: glance (model) |> select(r.squared)   Logistic Reg   MCR = $\frac{\#FP + \#FN}{\#preds}$
$MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$ $RMSE = \sqrt{MSE}$ lm(resp ~ poly(x = predictor, degree = 3, raw = TRUE), data = df) $\hat{p} = \frac{1}{1+e^{-z}}$ lm is linear model   mode glm(y~x, data = df, family = "binomial")
     predict(object = model, newdata = df)
ts-te & sample (x = c("tr","te"), size = n, replace = TRUE, probs = c(.8, .2))   FP: pred 1, true = 0
df & mutate(set_type = tr_te)   tr = df |> filter (set_type = "tr")   FN: pred 0, true = 1
     te = df |> filter(set_type = "te")