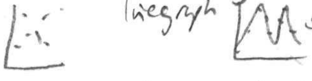
Villarin 6
EJPM193

(1) population – all sample – same, sampling error – error from sampling, precision, accuracy, bias – systematically under/overestimate
random sample – equal and independent, minimizes bias + makes possible to measure sampling error
samples of convenience – easily accessible researcher, volunteer bias – difference between volunteers
categorical – category/group, no magnitude on numerical scale, nominal = order, ordinal = no order
numerical – has magnitude and quantitative; continuous – any value, discrete – indivisible
predict response variable from explanatory variable (dependent; independent)
frequency – number of observations having particular value; freq. distribution – shows how often
experimental – assigns, observational – nature assigns
probability distributions – distribution of variable on whole population
pseudoreplication = design/sampling inappropriate for testing hypotheses – not independent

(2) displaying data
Bar graph, histogram – numerical variable, cumulative frequency distribution – fraction of observations ≤ that measurement
categorical contingency tables – categorical; grouped bar graph (height); categorical, mosaic (area) categorical

symmetric, asymmetric, skew, mode, bimodal
percentile = Xth percentile = X/100 quantile, eg 10th percentile = .1 quantile
Scatterplot – numerical; line graph – ordered series



Comparing histograms / cumulative frequencies

types	method
2 categorical	grouped bar graph, mosaic
numerical + categorical	grouped histograms, cum. freq. dist., boxplot
2 numerical	scatterplot, line plot, map

(3) describing data
Sample mean – average of measurements in sample; sample s.d. – spread of distribution
$$\text{variance} = \frac{\sum (Y_i - \bar{Y})^2}{n-1} \quad \text{(sum of squares)}$$
$$\text{mean } SD = \sqrt{\text{variance}} = \sqrt{\frac{\sum (Y_i^2) - n\bar{Y}^2}{n-1}}$$
proportion = fraction in category
$$\text{coefficient of variation} = \frac{SD}{\text{mean}} \times 100\%$$
IQR = 25% – 50% (med) – 75% (3rd quartile)

(4) uncertainty
population parameters vs sample estimates
μ, σ, p vs $\bar{Y}, s, \hat{p}$
sampling distribution = probability distribution of values for an estimate we might obtain when we sample
2SE = 95% of confidence interval
confidence intervals = range of values surrounding sample estimate

(5) probability
random trial = 2 or more outcomes where occurrence not predicted. sample space – all possible outcomes of trial; event = any subset of sample space
probability – proportion of random trials in which specified event occurs
probability distributions = list of all probabilities of all mutually exclusive outcomes of trials
venn diagrams; mutually exclusive – cannot occur simultaneously
discrete = each; continuous = probability density – area under
independent = occurrence of one does not affect other
Bayes' theorem: $Pr[A|B] = \frac{Pr[B|A]\,Pr[A]}{Pr[B]}$

(6) hypothesis testing
null hypothesis – nothing, alternative – change
two-sided/two-tailed – alternative hypothesis includes values on both sides of value specified by null; vs. one-sided
test statistic – quantity calculated from data used to evaluate how compatible the results are with that expected under null hypothesis
null distribution – sampling distribution of outcomes for test statistic under assumption null hypothesis is true – if null hypothesis were true
p-value: probability of obtaining data (or data showing ≥ null hypothesis)
$p \leq \alpha$ = reject
statistical significance level (α) = probability used as criterion for rejecting null hypothesis
α = probability of type I
power = probability random sample ⇒ reject false null hypothesis
↑ sample size = ↑ power = ↑ type I, ↓ type II
hypothesis = sufficient evidence?

decision	H_0 true	H_0 false
reject	type I	correct
no reject	correct	Type II

95% confidence interval ⇒ $\alpha = 0.05$

(7) proportions
binomial distribution – probability dist. for # of successes in indep. trials when prob. same in each trial; fixed #
adding probabilities; Agresti-Coull method
$$p' - Z\sqrt{\frac{p'(1-p')}{n+4}} < p < p' + Z\sqrt{\frac{p'(1-p')}{n+4}}$$
95%, Z = 1.96
99%, Z = 2.58
$$p' = \frac{X+2}{n+4}$$
Wald: $\hat{p} - Z\,SE_{\hat{p}} < p < \hat{p} + Z\,SE_{\hat{p}}$
$SE_{\hat{p}} =$
$$Pr[X \text{ successes}] = \binom{n}{X} p^X (1-p)^{n-X} = \left(\frac{n!}{X!(n-X)!}\right) p^X (1-p)^{n-X}$$
p = prob, X = # successes, n = trials

(8) X^2 goodness of fit – compares frequency data to model stated by null hypothesis. X^2 = measures discrepancy between observed + expected frequencies
critical value = value of test statistic that marks boundary of specified area in tail(s) of sampling distribution under H₀
assumptions: no categories has expected < 1, no more than 20% has expected freq < 5; use when too tedious to count sum of expected
Poisson: # of successes in blocks of time/space, when successes happen indep. of each other & occur w/ equal prob.

(9) Contingency analysis = estimates + tests for association between 2+ categorical variables
X^2 contingency test – same assumptions as
68.3, 95, 99.5
$Pr[\text{lower band} < Z < \text{upper band}] = Pr[Z > \text{lower band}] - Pr[Z > \text{upper band}]$

(10) Normal distribution – continuous probability dist describing bell curve.
standard normal distribution = normal dist w/ mean 0 and sd 1
Z = standard normal deviate = how many sd from mean.
Central Limit Theorem – sum or mean of large # of measurements randomly sampled from non-normal population approx. normally distributed.
If variable Y has normal dist in pop, the dist of means $\bar{Y}$ also normal.

(11) t-tests: assumptions: data are random sample from pop, variable is normally distributed
(12) 2-sample t-tests: paired – both treatments applied to every unit
2 sample – each treatment group composed of indep. random sample of units
paired – converted to single measurement, stating difference between them

$sd = \sqrt{\frac{\Sigma(Y_i - \bar{Y})^2}{n-1}}$ $\quad$ $se \doteq \sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} \Rightarrow SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$ $\quad$ binomial: $\hat{p} = \frac{X}{n}$ $\quad$ $SE_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/(n-1)}$

Agresti-Coull: $\left(p' - Z\sqrt{\frac{p'(1-p')}{n+4}}\right) < p < \left(p' + Z\sqrt{\frac{p'(1-p')}{n+4}}\right)$ $\cdot$ $p' = \frac{X+2}{n+4}$ $\quad Z = 1.96$ for 95% $\quad X$ = success, n = sample

$P = 2\left(\sum_{i=X}^{n} Pr[i \text{ successes}]\right)$ for $\frac{X}{n} > p_0$ or $P = 2\left(\sum_{i=0}^{X} Pr[i \text{ successes}]\right)$ for $\frac{X}{n} < p_0$ $\quad$ $\boxed{\text{poisson}}$: $\frac{e^{-\mu}\mu^X}{X!}$

$\boxed{X^2} = \sum \frac{(\text{observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$ $\quad$ df = # categories − 1 − # of parameters estimated from data $\quad$ μ = mean successes, X = successes

use $Pr[X]$

$\boxed{\text{Contingency}}$

odds $O = \frac{p}{1-p} = \hat{O} = \frac{\hat{p}}{1-\hat{p}}$ $\quad$ $OR = \frac{O_1}{O_2}$ = odds ratio $\quad$ $SE[\ln \hat{OR}] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$

$$X^2 = \sum_{\text{column}=1}^{c} \sum_{\text{row}=1}^{r} \frac{[\text{observed(column, row)} - \text{Expected(column, row)}]^2}{\text{Expected(column, row)}} \qquad df = (r-1)(c-1)$$

$=$ Expected (row i, column j) $= \frac{(\text{Row } i \text{ total})(\text{Column } j \text{ total})}{\text{Grand total}}$ $\quad$ $\ln(\hat{OR}) - 2SE[\ln(\hat{OR})] < \ln(OR) < \ln(\hat{OR}) + 2SE[\ln(\hat{OR})]$

$\boxed{\text{normal}}$ $f(Y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(Y-\mu)^2}{2\sigma^2}}$ $\quad$ $Z = \frac{\overset{\text{value}}{Y} - \mu}{\sigma} = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}}$ $\quad \sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$

$n\uparrow$, binomial prob. dist $\Rightarrow$ mean np, $sd = \sqrt{np(1-p)}$

normal approx to binomial dist: $Pr[X \geq \text{observed}] = Pr\left[Z > \frac{\text{observed} - np}{\sqrt{np(1-p)}}\right]$

$np, n(1-p) \geq 5$ $\quad$ Continuity correction for normal approx to binom. dist

$Pr[X \geq \text{observed}] = Pr\left[Z > \frac{\text{observed} - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right]$ $\quad$ $Pr[X \leq \text{observed}] = Pr\left[Z < \frac{\text{observed} + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right]$

(11) $\boxed{t\text{-test}}$ −1 sample $\quad Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}}$ $\quad SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$ $\quad t = \frac{\bar{Y} - \mu}{SE_{\bar{Y}}} = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}$ $\quad df = n-1$

$-t_{0.05(2),df} < \frac{\bar{Y} - \mu}{SE_{\bar{Y}}} < t_{0.05(2),df} \Rightarrow \bar{Y} - t_{0.05(2),df} SE_{\bar{Y}} < \mu < \bar{Y} + t_{0.05(2),df} SE_{\bar{Y}}$

$\bar{Y} - t_{\alpha(2),df} SE_{\bar{Y}} < \mu < \bar{Y} + t_{\alpha(2),df} SE_{\bar{Y}}$ $\quad$ $P = Pr[t < y] + Pr[t > y]$

$X^2 = (n-1)\frac{s^2}{\sigma^2}$ $\quad$ $\frac{df\, s^2}{X^2_{\frac{\alpha}{2},df}} < \sigma^2 < \frac{df\, s^2}{X^2_{1-\frac{\alpha}{2},df}}$ $\Rightarrow$ Confidence interval for variance, or not normal

estimate $= s^2$
parameter $= \sigma^2$

(12) $\boxed{\text{2-sample }t\text{-test}}$ paired − test null hypothesis that the mean diff of paired measurements = specified value. make sure to check 1 vs. 2 tails!

$t = \frac{\bar{d} - \mu_{d0}}{SE_{\bar{d}}}$ $\quad P = Pr[t_y < -t] + Pr[t_y > t] = 2Pr[t > t]$ $\quad$ $\bar{d} - t_{\alpha(2),df} SE_{\bar{d}} < \mu_d < \bar{d} + t_{\alpha(2),df} SE_{\bar{d}}$

assumptions: sampling units randomly sampled, paired diff have normal distr in population

2 sample $\quad SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ $\quad s_p^2 = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2}$ $\quad df_1 = n_1 - 1$, $df_2 = n_2 - 1$ $\quad t = \frac{(\bar{Y}_1 - \bar{Y}_2)}{SE_{\bar{Y}_1 - \bar{Y}_2}}$ $\quad SE = \frac{s_d}{\sqrt{n}}$

s_p: pooled sample variance

$df = df_1 + df_2 = n_1 + n_2 - 2$ $\quad$ if grouped, use this, not X^2

assumptions: each of the two samples is a random sample
the numerical variable is normally distributed
sd + variance is same in both pop. else → welch's t-test

F-test evaluates whether two pop. variances are equal
$H_0 = \sigma_1^2 = \sigma_2^2$, $H_A\ \sigma_1^2 \neq \sigma_2^2$, $F = \frac{s_1^2}{s_2^2}$
for testing variances $\quad$ Levene's = roughly symmetrical within all groups, ≥2 groups

binomial − fixed-trial limit $\quad$ geometric − open-ended limit $\quad$ poisson − how many times this thing, within a certain range

Ch 19

Simulation — Use computer to create 1) sample pop w/ null 2) calculate test statistic 3) repeat 4) make null distribution 5) compare test statistic

Randomization — categorical — tops one, randomized else 1) randomize 2) calculate association 3) repeat

Bootstrap — resampling to compute sampling distribution 1) random from orig. size 2) estimate 3) repeat 4) sample distribution = std error = std dev

Ch 18

F measures improvement in fit of model to data with term describing explanatory variable included

Ch 13 transformations — arcsin: proportions; sqrt: counts; square: left; antilog: left; reciprocal: right

Mann-Whitney U-test — lowkis $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$, $U_2 = n_1 n_2 - U_1$ — p-values

Ch 14

blinding, replication, controls, covariates, control group, randomization, balance, blocking, sample size: $n = \frac{4p(1-p)}{\text{uncertainty}^2}$ prop.

odds — $n = \frac{4}{\text{uncertainty}^2}\left(\frac{1}{p_1} + \frac{1}{1-p_1} + \frac{1}{p_2} + \frac{1}{1-p_2}\right)$ diff: $n = \frac{8\sigma^2}{\text{uncertainty}}$ mean: $n = 4\left(\frac{\sigma^2}{\text{uncertainty}}\right)^{2}$ $\frac{1}{2}$ interval

n = sample size

k = # of groups

Ch 15 ANOVA	SS	df	mean squares	F ratio
groups	$SS = \sum n_i(\bar{Y}_i - \bar{Y})^2$	$k-1$	$\frac{SS_{groups}}{df_{groups}}$	$\frac{MS_{groups}}{MS_{error}}$
error	$SS = \sum s_i^2(n_i - 1)$	$N-k$	$\frac{SS_{error}}{df_{error}}$	

$R^2 = \frac{SS_{groups}}{SS_{total}}$

Ch 16:

$SS_{prod} = \sum(X-\bar{X})(Y-\bar{Y}) = \sum(XY) - \frac{\sum X \sum Y}{n}$

$SS_{squares} = \sum(X-\bar{X})^2 = \sum(X^2) - \frac{(\sum X)^2}{n}$

$\text{Covariance}(X,Y) = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{n-1}$

correl. coefficient = $r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}}$ $df = n-2$

bivariate normal

pop correlation — random sample $z - Z_{crit}\sigma_z < \zeta < z + Z_{crit}\sigma_z$

$z = 0.5\ln\left(\frac{1+r}{1-r}\right)$ $\sigma = \sqrt{\frac{1}{n-3}}$, $r = \frac{e^{2z}-1}{e^{2z}+1}$

t-test of linear correl. : $n-2$ df, $t = \frac{r}{SE_r}$ $SE_r = \sqrt{\frac{1-r^2}{n-2}}$

Ch 17

regression slope : β ; β : $b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$ $SE_b = \sqrt{\frac{MS_{resid}}{\sum(X_i-\bar{X})^2}}$ $MS_{resid} = \frac{\sum(Y_i - \bar{Y})^2 - b\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n-2}$

t-test for slope $t = \frac{b - \beta_0}{SE_b}$

ANOVA	SS	df	mean squares	F ratio
regression	$\sum(\hat{Y}_i - \bar{Y})^2$	1	$\frac{SS_{regress}}{df}$	$\frac{MS_{regress}}{MS_{resid}}$
residual	$\sum(Y_i - \hat{Y}_i)^2$	$n-2$	$\frac{SS_{resid}}{df_{resid}}$	

$R^2 = \frac{SS_{regression}}{SS_{total}}$

Likelihood

$\binom{n}{k} p^k (1-p)^{n-k}$

$\frac{\chi^2_{1,0.05}}{2} = 1.92$

Op. [illegible] like

$2\ln\left(\frac{\text{Likelihood}}{\text{Likelihood}}\right)$