Experimental Stats

- **Sampling Bias**: self-selection, non-response, undercoverage, survivorship **Confounder**: correlated/causal /w indepen var & causal /w depen var
- **Probability Sampling**: random, stratified, clustered **Non-Probability Sampling**: voluntary, convenience, snowball

Variability & Correlation vs Covariance

- **Variance**: Average squared distance from the mean, $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$ **Z-Score**: $z = \frac{x_i - \bar{x}}{\sigma}$
- **Covariance**: $(-\infty, \infty)$ scaled by variances, measures linear relationship between two variables, $\mathrm{Cov}(X,Y) = \sum(x_i-\bar{x})(y_i-\bar{y})$
- **Correlation**: $[-1,1]$ unitless, normalized covariance, $\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$

Intro Probability

- **Combination**: Unordered sampling without replacement: $\binom{n}{r} = \frac{n!}{r!(n-r)!}$ **Permutation**: Ordered sampling without replacement: $\frac{n!}{(n-r)!}$
  - **Ex**: $P(3 \text{ spades}, 2 \text{ hearts}) = \frac{({}_{13}C_3)({}_{13}C_2)}{{}_{52}C_5}$, Split 12 people into 3 groups of 4: $\frac{\binom{12}{4}\binom{8}{4}\binom{4}{4}}{3!}$
- **Unordered Sampling with Replacement**: $\binom{n+r-1}{r}$ **Ordered Sampling with Replacement**: n^r
- **Union**: $P(A\cup B) = P(A) + P(B) - P(A\cap B)$ **Complement**: P(X occurs at least Y times) = 1 - P(X doesn't occur at least Y times)

Conditional Probability

- **Bayes Formula**: $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A\cap B)}{P(B)}$ **Chain Rule**: $P(A\cap B) = P(B|A)\cdot P(A) = P(A|B)\cdot P(B)$
- **Independence**: A and B independent if: 1. $P(A|B) = P(A)$, or 2. $P(B|A) = P(B)$, or 3. $P(A\cap B) = P(A)\cdot P(B)$
- If A and B are independent so are pairs $(A^c, B),\ (A, B^c),\ (A^c, B^c)$
- **Ex**: Healthy or Sick patient can test (+) or (-) then $P(+) = P(+|H)\cdot P(H) + P(+|S)\cdot P(S)$

Bayesian vs Frequentist

- **Probability**: Chance of event/data (e.g. training/experiment data) occurring among a set of MECE events, sum of possible probas = 1
- **Likelihood**: Chance of a hypothesis/params (e.g. model/distribution) occurring among set of infinite non-MECE hypotheses, sum of likelihoods ≠ 1
- **Proba Func**: Gives probas of varying data (events) given set of params **Likelihood Func**: Gives relative likelihoods of varying params given data
- **Frequentist**: Probability measured as relative freq of event over long run, assume data $\boldsymbol{X}$ is random → find fixed params $\vec{\theta}$ that generated $\boldsymbol{X}$
  - **MLE**: Max likelihood: $\vec{\theta}^*$ = argmax $P(\boldsymbol{X}|\vec{\theta})$ = argmin $(-\log P(\boldsymbol{X}|\vec{\theta}))$, ML: setup model as MLE optimization → use GD e.g. logreg /w log loss
- **Bayesian**: Probability measured as confidence in occurrence of event, assume data $\boldsymbol{X}$ is fixed → find param $\vec{\theta}$ prob distribution given $\boldsymbol{X}$ and prior
  - **Posterior**: (distribution) $P(\vec{\theta}|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|\vec{\theta})\cdot P(\vec{\theta})}{P(\boldsymbol{X})}$ where $P(\boldsymbol{X}|\vec{\theta})$ = likelihood func & $P(\vec{\theta})$ = prior (distributions), $P(\boldsymbol{X})$ = evidence (fixed)
  - **MAP**: Max posterior: $\vec{\theta}^*$ = argmax $P(\vec{\theta}|\boldsymbol{X})$ = argmax $P(\boldsymbol{X}|\vec{\theta})\,P(\vec{\theta})$, update prior := posterior when new data, MAP = MLE if prior is uniform
  - Bayesian methods useful when 1. prior knowledge, 2. small dataset (e.g. election, market crash), 3. online learning from one sample at a time

Moments

- **Mean**: $\mu = E(X) = \sum x_i\cdot p(x_i)$ **Var**: $\sigma^2 = \mathrm{Var}(X) = E[X^2] - \big(E[X]\big)^2 = \sum(x_i-\mu)^2\cdot p(x_i)$ **Skew**: $\frac{E\left[(X-\mu)^3\right]}{\sigma^3}$ **Kurtosis**: $\frac{E\left[(X-\mu)^4\right]}{\sigma^4}$

Sampling Distribution

- **Sampling Distribution of Sample Means**: $\bar{X}$ related to pop mean by: $\mu_{\bar{X}} = \mu$ and to pop stdev by: $\sigma_{\bar{X}} = \text{standard error of } \bar{X} = \frac{\sigma}{\sqrt{n}}$
- **Central Limit Thm**: If $n \geq 30$, then regardless of population shape, $\bar{X}$ has $\bar{X} \sim N(\mu_{\bar{X}}, \sigma_{\bar{X}}) = N(\mu, \frac{\sigma}{\sqrt{n}})$ distribution

Discrete Distributions - PMF $p(X=r)$, CDF $p(X\leq r)$

- **Bernoulli** $X \sim \mathrm{Bern}(p)$ - experiment with success prob p, failure prob $1-p$ (e.g. roll 6 or not)
  - Two events: Fail $P(X=0) = 1-p$, Success $P(X=1) = p$ Moments: $E[X] = p,\ \mathrm{Var}(X) = p(1-p)$
- **Binomial** $X \sim \mathrm{Bin}(n,p)$ - # of successes in n Bernoulli trials (e.g. # of 6's rolled in 5 rolls)
  - $\mathrm{PMF}(n,p) = P(X=r) = \binom{n}{r}p^r(1-p)^{n-r}$ where $\binom{n}{r}$ = # of subsets, $p^r(1-p)^{n-r}$ = r success + n-r failures in subset
  - $E[X] = np,\ \mathrm{Var}(X) = np(1-p)$
- **Poisson** $X \sim \mathrm{P}(\lambda)$ - # of successes in infinite Bernoulli trials (e.g. # of earthquakes in a year, # of typos in a book)
  - $E[X] = \lambda,\ \mathrm{Var}(X) = \lambda,$ If $P(\lambda = 2 \text{ hurricanes/yr on avg})$ and want # of hurricanes per 3 years, then use $P(\lambda = 2\cdot 3 = 6)$

- **Geometric** $X \sim \text{Geom}(p)$ - # of Bernoulli trials until & including first success (e.g. # of coin flips to get first heads)
  - $\text{PMF} = P(X = k) = (1-p)^{k-1}p,$ k-1 failures before first success + success, $\text{CDF} = P(X \leq k) = 1-(1-p)^k$
  - $E[X] = \frac{1}{p},\ \text{Var}(X) = \frac{1-p}{p^2},$ Memoryless: The prob of s more trials until first success given k failures already, is the same as taking s trials

<u>Continuous Distributions</u> - Integrated PDF $p(a \leq X \leq b) = \int_b^a f(x)\,dx,$ CDF $P(X \leq x) = \int_{-\infty}^x f(t)\,dt$

- **Uniform** $X \sim \text{U}[a,b]$ - choose randomly equally likely point in $[a,b]$ (e.g. min postman arrives in next hour)
  - $\text{PDF} = f(x) = \frac{1}{b-a}, \text{for } a \leq x \leq b$ take the integral of this from a to b
  - $E[X] = \frac{a+b}{2},\ \text{Var}(X) = \frac{(a-b)^2}{12},$ Universality of Uniform: sample from any distrib X, feed samples to the CDF of X to sim $\text{U}[0,1]$ distrib
- **Normal** $\text{N}(\mu,\sigma)$ - (e.g. heights, darts on dartboard, IQ)
  - $E[X] = \mu,\ \text{Var}(X) = \sigma^2$
- **Exponential** $X \sim \text{Exp}(\lambda)$ - time until next success, cts geometric equivalent (e.g. time until next accident, time between two Poisson events)
- **Gamma** $X \sim \text{Gamma}(n, \frac{1}{\lambda})$ - time until nth occurrence, generalized exponential rv (e.g. time until 3rd accident)
- **Chi-Squared** $X \sim \chi^2(k)$ - sum of k samples squared from std normal distribution

<u>Errors & Successes</u>

- $\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$ $= P(\text{reject } H_0 | H_0 \text{ true})$ $= \alpha$ = Significance Level = Type 1 Error Rate
- $\text{FNR} = \frac{\text{FN}}{\text{FN+TP}}$ $= P(\text{accept } H_0 | H_0 \text{ false})$ $= \beta$ = Type 2 Error Rate
- $\text{TNR} = \frac{\text{TN}}{\text{TN+FP}}$ $= P(\text{accept } H_0 | H_0 \text{ true})$ $= 1 - P(\text{reject } H_0 | H_0 \text{ true})$ $= 1-\alpha$ = Specificity
- $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$ $= P(\text{reject } H_0 | H_0 \text{ false})$ $= 1 - P(\text{accept } H_0 | H_0 \text{ false})$ $= 1-\beta$ = Sensitivity = Recall = Power (π)
- **Power Analysis** = Estimate one of the 4 BEAN variables given the other 3 (e.g. find n given $\alpha = 0.05, 1-\beta = 0.80$, and some min effect size)
  1. <u>Beta β</u>: If β↓ we have a smaller tolerance for false negatives, if $(1-\beta)$↑ then π↑ then n↑
  2. <u>Effect Size</u>: Distance between sampling distribution means, if ES↑ then π↑ (note: n doesn't affect ES)
  3. <u>Alpha α</u>: Set by us, corresponding critical value splits β and α, if α↑ then n↓ and we have a larger tolerance for false positives
  4. <u>n Sample Size</u>: If n↓ then SE $\frac{\sigma}{\sqrt{n}}$↓ and H_a distribution tightens and π↑
- **Overpowered**: n too large, classify tiny differences as significant **Underpowered**: n too small, unable to detect important differences

<u>Confidence Interval</u>

- **CI Formula**: Point Estimate ± Error Bound = Point Estimate ± (Critical Value * Standard Error) **CI Size**: 1-α=CC↑ or n↓ then CI width ↑
- **CI for μ, When σ is Unknown**: $\bar{x} \pm t_{\alpha/2,\, n-1} \frac{s}{\sqrt{n}}$ **CI for p Population Proportion**: $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$

<u>Hypothesis Testing</u>

- **Test Stat Formula**: $\frac{\text{statistic}-\text{parameter}}{\text{stdev of the statistic}} = \frac{\text{statistic}-\text{parameter}}{\text{standard error}}$
- **p-Value**: Converts test stat into conditional prob that given H_0, we observe a value at least as extreme as this one, if $p < \alpha$ reject H_0
- **Effect Size**: Measures magnitude of experimental effect, p-val says intervention works, effect size says how much the intervention works
- **Right Tailed Test**: One-sample $H_0 : \mu = \mu_0 \text{ vs } H_a :\ \mu > \mu_0$, Two-sample $H_0 : \mu_1 - \mu_2 = 0 \text{ vs } H_a :\ \mu_1 - \mu_2 > 0$
- **Non-Parametric Test**: Tests that don't assume population distribution is normal (e.g. chi-squared, wilcoxon signed rank)
- **Z vs T-Test**: Use Z when σ known or n large or proportion test, else use T-test since T-distribution thicker tails compensate for smaller n
- **One Sample Mean Tests** - (e.g. are soda bottles filled above 20oz?)
  - One Sample Mean T-Test T-Stat: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$, compare with critical value $\pm t_{\alpha,\, n-1}$ for single tail tests, or $\pm t_{\alpha/2,\, n-1}$ for two-tail tests
  - Cohen's D Effect Size: $d = \left|\frac{\bar{x}-\mu_0}{s}\right|$ where d measures num of stdevs between the two means $d = .2$ is small effect, $d = .8$ is large
  - One-Sample Paired Mean T-Test: Test mean difference of one population twice sampled (e.g. does SAT prep raise SAT score?)
- **Two-Sample Mean Tests** - (e.g. compare avg revenue per user or avg posts per user between control and treatment)
  - Z-Test, T-Test with Uknown but Equal Stdevs, T-Test with Unknown Unequal Stdevs (Aspin-Welch's Test)
- **Proportion Z-Tests**
  - One-Sample Proportion Z-Stat: $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0 q_0}{n}}}$, (e.g. are scanners erroring more than 2% of time?)

- Two-Sample Proportion Z-Stat: $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(\frac{1}{n_1} + \frac{1}{n_2})}}$, (e.g. compare CTR between control and treatment)

- **Chi-Squared Distribution Test**
  - Goodness of Fit Test - Check distribution of categorical var matches expected distribution (e.g. each choice on a MC exam equally likely)
  - Independence Test - Check if two categorical vars are related or not based on distributions (e.g. do RHers like same subjects as LHers)
- **ANOVA Mean Test** - omnibus test, tells us at least one of the means are significantly different, but not which one
  - Compare variation within each group with variation between each group, if between groups variation >> within, one or more groups stand out
  - Use ANOVA for 3+ groups, since T-tests have $\alpha = .05$ FPR, 3 groups have $1 - (.95)^3 = 14.3\%$ chance of at least one FP
  - If H_0 rejected, run post-hoc to see which mean differed (e.g. Dunnett), compare each pair of groups using lower α (thus lower π)
  - One-Way ANOVA - check if any groups from independent categorical var differ in the dependent var (e.g. black, green, or no tea on alertness)
  - Two-Way ANOVA - check if interaction between two independent categorical vars on the dependent var (e.g. gender + stress level affect job?)