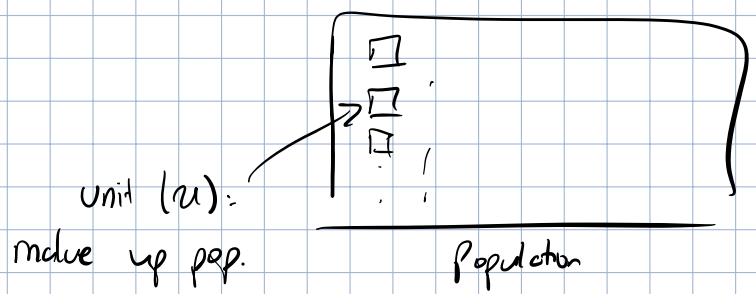
POPULATION

Defn: finite set of elements, like entire universe

Notation: P

Ex: //

1. Population of world
2. All posts on Insta



Variable: attribute of unit

↳ Notation: x_u → attr. x of unit u

Population attribute: property of P, summaries

↳ Notation: $a(P)$

Ex: //

1. avg. income for all humans on Earth
2. total # of likes on Insta posts.

Population attributes

Pop. attributes are fn:

$$a(P) = f(y_1, \ldots, y_N)$$

Ex: //

1. Pop. total: $a(P) = \sum_{i=1}^{N} y_i$
2. Count: $a(P) = \sum_{u \in P} I_A(y_u)$, $I_A = \begin{cases} 1, & y \in A \\ 0, & y \notin A \end{cases}$

Numerical / graphical

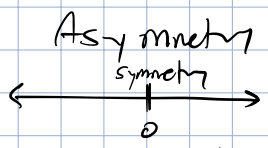
Types of attr.

Location

Defn: describes centre of all variates

- Avg.
- Median

Skewness



- Pearson's moment of coef.

Spread

Defn: variability of variate.

- Population variance:

$$a(P) = \frac{1}{N} \sum (y_i - \bar{y})^2$$

also $N-1$

- S.D.
- Coefficient of variation

$$a(P) = \frac{SD_P(y)}{\bar{y}}$$

Order of attributes can yield insights

↳ Notation: $y_{(1)} \leq \dots \leq y_{(N)} \Rightarrow y_{(i)}$ is i^{th} smallest variate in pop.

↳ Location Attribute: min, max, median, quartiles ...

↳ Variability " : range, IQR, Median abs. deviation

range: $y_{(N)} - y_{(1)}$

$$a(P) = \underset{u \in P}{\text{median}} \left(y_u - \underset{v \in P}{\text{median}}\, y_v \right)$$

median of diff. b/w median & data points.

Attribute Properties

Q: What happens if variates change? Will attribute also change?

2 important terms:

① Invariance: no change

② Equivariance: commesurate change

LOCATION:

↳ Invariance:

$$a(y_1 + b, \dots, y_N + b) = a(y_1, \dots, y_N)$$

↳ Equivariance:

$$a(y_1 + b, \dots, y_N + b) = a(y_1, \dots, y_N) + b$$

Ex:// Show pop. avg. is location equivariant.

① Define original $a(P)$

$$a(P) = \frac{1}{N}\sum_{i=1}^{N} y_i$$

② Change the units

$$a(P^*) = \frac{1}{N}\sum_{i=1}^{N}(y_i + b)$$

P^*: $\{y_1 + b, \ldots, y_N + b\}$

② Put RHS in form of original attr.

$$a(P^*) = \frac{1}{N}\sum_{i=1}^{N} y_i + \frac{1}{N} \cdot Nb$$

$$= a(P) + b \quad \blacksquare$$

SCALE:

↳ Invariance:

$$a(my_1, \ldots, my_N) = a(y_1, \ldots, y_N)$$

↳ Equivariant

$$a(my_1, \ldots, my_N) = m \cdot a(y_1, \ldots, y_N)$$

LOCATION - SCALE

↳ Invariance

$$a(my_1 + b_1 \ldots, my_N + b) = a(y_1 \ldots y_N)$$

↳ Equivariant:

$$a(my_1 + b_1 \ldots, my_N + b) = m \cdot a(y_1 \ldots y_N) + b$$

Ex:// Show pop. avg. is location-scale invariant.

$$a(m \cdot y_1 + b, \ldots) = \frac{1}{N}\sum_{i=1}^{N}(my_i + b)$$

$$= \frac{m}{N}\left(\sum_{i=1}^{N} y_i\right) + b$$

$$= m \cdot a(P) + b \quad \blacksquare$$

REPLICATION

Duplicate pop various k times $\rightarrow$ new pop.

$$P^k = \{ \underbrace{y_1 \dots y_1}_{k}, \dots, \underbrace{y_N \dots y_N}_{k \text{ copies}} \}$$

→ Invariance:

$$a(P^k) = a(P)$$

→ Equivariance:

$$c(P^k) = k \cdot a(P)$$

Ex:// Show pop. avg is replication invariant

$$a(P^k) = \frac{1}{NK} \sum_{i=1}^{NK} y_i$$

} diff. way of writing sum

$$= \frac{1}{NK} \sum_{i=1}^{N} k \cdot y_i$$

$$= \frac{1}{N} \sum_{i=1}^{N} y_i$$

$$= a(P)$$

Influence, Sensitivity Curves, Breakdown Points

Influence: $\Delta(a, u) = a(1, \dots, y_{u-1}, y_u, y_{u+1}, \dots, y_n)$ ← w/ unit u

$$- a(1, \dots, y_{u-1}, y_{u+1}, \dots y_n)$$ ← w/out unit u

- If some unit u has high infl. compared to other units:
  - ① Error point
  - ② Interesting point

Ex:// $a(P)$ = mean.

w/ u: $\bar{Y}$

w/out u: $\frac{1}{N-1} \sum_{\substack{l \in P \\ l \neq u}} y_l = \frac{N\bar{Y} - Y_u}{N-1}$

$$\Delta(a, u) = \bar{Y} - \frac{N\bar{Y} - Y_u}{N-1} = \boxed{\frac{y_u - \bar{Y}}{N-1}}$$

In R: simulate diff. u and find Δ for all → plot

Sensitivity Curves:

Q: how sensitive is attribute to data addition?

Defn:

$$SC(a, y) = N\left[a(y_1, \ldots, y_{n-1}, y) - a(y_1, \ldots, y_{n-1})\right]$$

↑ change y in range $(f(y))$

Ex.// Sensitivity curve for avg.

① Define pop:

$$P = \{y_1, \ldots, y_{n-1}\} \qquad P^* = \{y_1, \ldots, y_{n-1}, y\}$$

② SC:

$$sc(a, y) = N\left[a(P^*) - a(P)\right]$$

③ Solve for $a(P^*)$, $a(P)$

$$a(P^*) = \frac{1}{N}\left(\sum_{i=1}^{N-1} y_i + y\right) = \frac{(n-1)\,\bar{y}_{n-1} + y}{N}$$

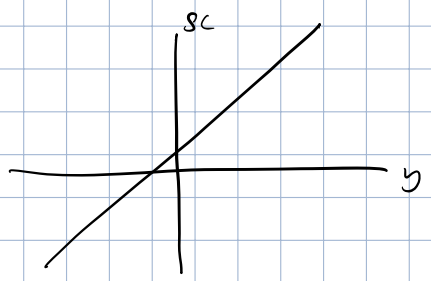
$$a(P) = \frac{1}{N-1}\sum_{i=1}^{N-1} y_i = \bar{y}_{N-1}$$

④ Plug:

$$sc(a, y) = N \cdot \left(\frac{(n-1)\,\bar{y}_{n-1} + y}{N} - \bar{y}_{N-1}\right)$$

$$= y - \bar{y}_{n-1}$$

⑤ Plot:



⑥ Interpret:

Avg. is very sensitive to extreme values

Other examples:

① S.C. for median
② " 2nd quartile
③ " maximum

} Review notes

Breakdown points:

Q: how large of pop. in data must be bad for attr. to break?

data: $-\infty, \infty$

error (near – any) = Inf.

For mean: 1 datapoint $\Rightarrow 1/N$

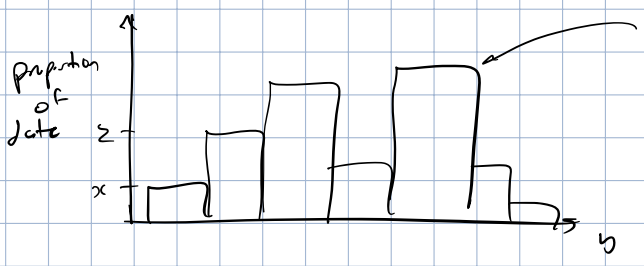
For median: $N/2$

If prop. high $\Rightarrow$ attr. is robust to bad data.

Graphical Attributes

Visual repr. of pop is an attribute $\leftarrow$ summarizes pop.

Histogram:

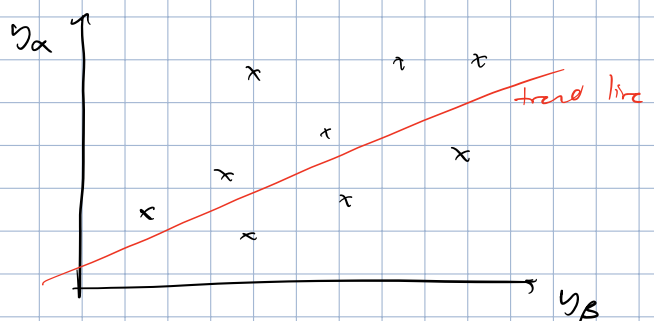


Bucket defn:

A: equal sized bins

B: same # of datapoints
↳ bins width changes (lot of datapoints → width ↓)

Scatter plot:



Might need to change opacity / add jitter if multiple datapoints w same coord.

Power transformations

Defn:

$$T_\alpha(y) = \begin{cases} y^\alpha, & \alpha > 0 \\ \log(y) & \alpha = 0 \\ -(y^\alpha), & \alpha < 0 \end{cases}$$

Monotonic transformation:

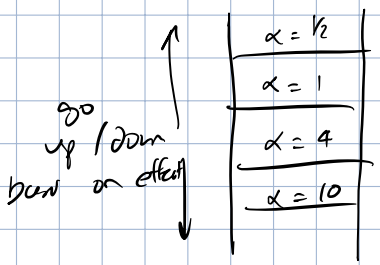
$$y_a < y_b \Rightarrow T_\alpha(y_a) < T_\alpha(y_b)$$

Why transform?

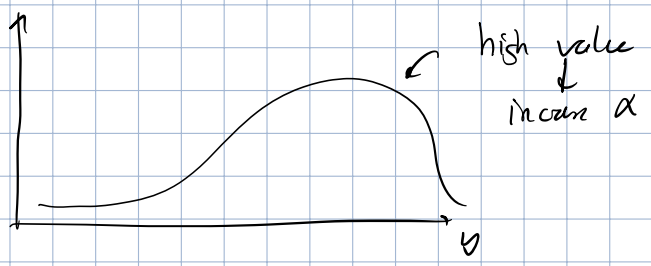
A: More symmetric histograms

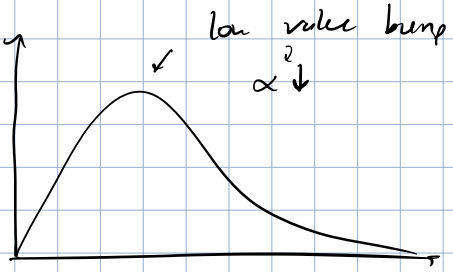
B: Linearize scatterplots

Tukey ladder:



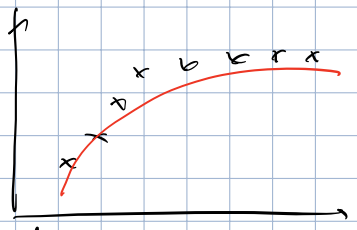
Bump rule #1: more symmetric histogram.



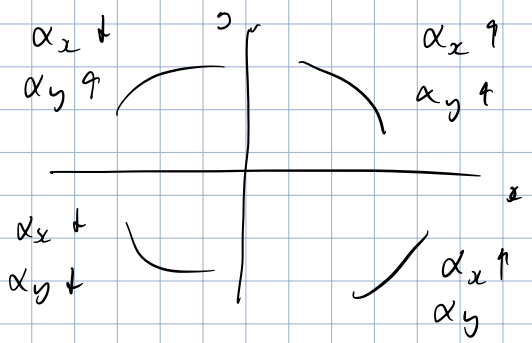


Bump rule #2: linear scatterplot

1. Find the "bump" in your graph



2. Use Tukey's ladder



Order, Rank, Quantiles

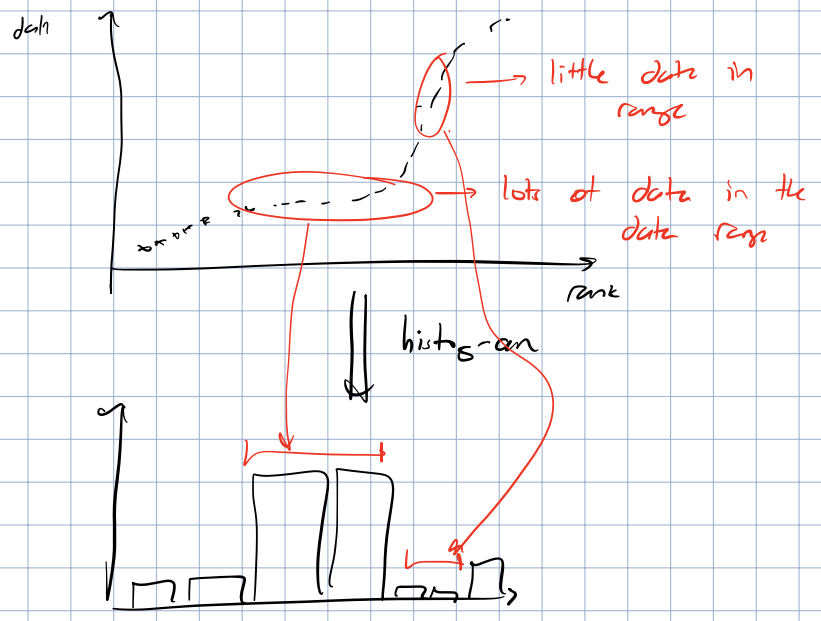
Rank:

If y_i is $y_{(k)}$ (ex. k^{th} smallest), then $r_i = k$

Ex://

y_i:	100	-50	1024	6
r_i:	3	1	4	2

Scatterplot of rank vs. data



Quantiles:

$$Q_y(p) \Rightarrow \text{datapoint w/ rank } p \cdot N$$

Ex: $Q_y(1/2) \Rightarrow$ datapoint w/ rank $N/2 \Rightarrow$ median

$Q_y(0.3) \Rightarrow 30^{th}$ percentile (30% of tot < datapoint)

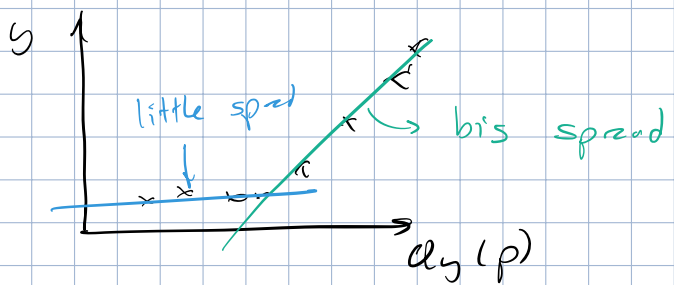
Measure of location

Spread quantiles:

IQR: $Q_y(0.75) - Q_y(0.25)$

Range: $Q_y(1) - Q_y(1/N)$

Slope of 2 quantiles $\propto$ spread of data



Implicit Attributes

Defn: attr. that summarize pop., but usually an optimization soln.

Minimum of function: $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}}\ \rho(\theta; P)$ ← optimization func.

Θ: list of all possible θ

$$= \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{u \in P} \rho(\theta; u)$$

argmin can also be used for max:

$$\hat{\theta}_{max} = \underset{\theta \in \Theta}{\operatorname{argmin}} -\rho(\theta; P)$$

Ex:

Least squares: $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{u \in P} (y_u - \theta)^2$

$$\vdots$$

$$= \bar{y}$$

Least absolute deviation: $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{u \in P} |y_u - \theta|$

$$\vdots$$

$$= \operatorname{med}(y)$$

θ can be vector valued!

Linear regression:

$$\hat{\theta} = (\hat{\alpha}, \hat{\beta}) = \underset{(\alpha, \beta) \in \mathbb{R}^2}{\operatorname{argmin}} \sum_{u \in P} [y_u - (\alpha + \beta x_u)]^2$$

↳ Influence of datapoint still applicable in implicit attr.

$$\Delta(\vec{\theta}, u) = \|\vec{\theta} - \vec{\theta}_{[-u]}\|_2 \Rightarrow \text{Euclidean norm}$$

Robust Regression

Methods of dealing w/ influential datapoints?

① Don't do anything

② Remove datapoints

↳ Drastic change to line of best fit

↳ Line doesn't summarize all datapoints

③ Weight lager influens smaller (weighted liner regression)
↳ Manual tuning

④ Robust regression: weight based off residuals (high residuals → low weight)

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{u \in P} \rho(r_u)$$

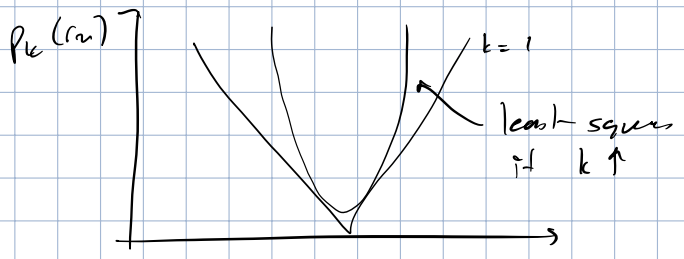
Liner reg: $\rho(r_u) = r_u^2$

Weighted ll: $\rho(r_u) = w_u \cdot r_u^2$

Huber loss:

$$\rho_k(r_u) = \begin{cases} \frac{r_u^2}{2}, & |r_u| \le k \\ k|r_u| - \frac{k^2}{2}, & |r_u| > k \end{cases}$$

⇒ lower weight for high residuals



How to choose k?

Theoretical: $k = 1.3$

Common choice: $k = 1, 2, \ldots$

Con: no close form soln for θ

Gradient Descent

Objective: $\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{u \in P} \rho(\theta)$

Also:

1. Initialize $i \leftarrow 0$, θ_0
2. Loop:
   i) Gradient:

$$s_i = \nabla \rho(\theta; \rho)\Big|_{\theta_i} = \left[\begin{array}{c} \frac{\partial \rho}{\partial \theta_1} \\ \vdots \\ \frac{\partial \rho}{\partial \theta_n} \end{array}\right]\Bigg|_{\theta = \theta_i}$$

   ii) Normalize:

$$d_i = \frac{s_i}{\|s_i\|}$$

   iii) Line search:

$$\hat{\lambda}_i = \underset{\lambda > 0}{\operatorname{argmin}} \; p(\theta_i - \lambda d_i)$$

iv) Make step:

$$\theta_{i+1} = \theta_i - \lambda d_i$$

step part descent

v) Check converge:

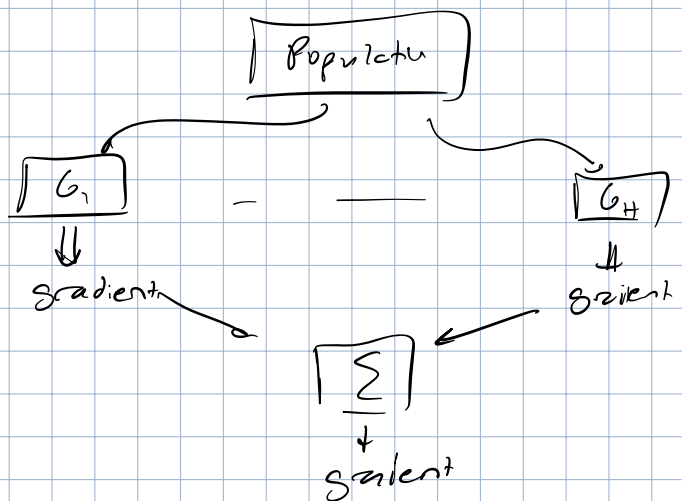
If θ is converging $\Rightarrow$ return θ_{i+1} ← L_1 norm of θ_{i+1} & θ_i is $< \varepsilon$

Else: $i \leftarrow i+1$, loop

Doesn't always get global minima

Batch gradient descent

Batch gradient calculations b/c not dependent on each other



Batch gradient descent
⇓
Goes through all data b/f updating param

Another optimization: calculate gradient off sample of pop. & use fixed step size.

Batch sequential:

1. Divide pop. into batches
2. Run G.D. on batch
3. Update θ
4. Move to next batch

} until θ converg.

Batch stochastic:

1. Randomly sample pop.
2. Run G.D. on sample

3. Update θ

4. Repeat 1 → 4 until converge.

Batch also take more steps, but less time/step

Systems of Equations

Alt. defn of implicit attribute: θ that solves

$$\Psi(\theta; P) = \sum_{u \in P} \Psi_u(\theta; u) = \vec{0}$$

sys. of equations

Ex:// 1.
$$\Psi(\theta; P) = \sum_{u \in P} \Psi_u(\theta; u) = \sum_{u \in P} w_u(y_u - \theta)$$

θ that best solves ⇒ weighted avg.

2. Weighted least squares problem:

$$\Psi(\theta; P) = \sum_{u \in P} \Psi_u(\theta; u) = \sum_{u \in P} \rho_k'(y_u - \alpha - \beta(x_u - c)) \begin{pmatrix} 1 \\ x_u - c \end{pmatrix} = \vec{0}$$

Q: How do we find θ that solves sys. of equations?

A: Root finding methods!

① Newton's Method

1. $i \leftarrow 0$
2. Loop:
   a. Update: $\hat{\theta}_{i+1} = \hat{\theta}_i - \frac{\Psi(\hat{\theta}_i; P)}{\Psi'(\hat{\theta}_i; P)}$
   b. Check convergence: $\hat{\theta}_i \approx \hat{\theta}_{i+1}$?
      - Yes: return $\hat{\theta}_{i+1}$
      - No: Loop
3. Return $\hat{\theta}_i$

Iteratively Reweighted Least Squares

Obj: Find $\hat{\theta} = (\alpha, \beta)$ that solves:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^2}{\operatorname{argmin}} \sum_{u \in P} \rho(\overbrace{y_u - \alpha - \beta(x_u - c)}^{\text{residual } r_u})$$

Solving derivative:

$$g = \nabla \rho(\theta;\ P) = \sum_{u \in P} \rho'(r_u) \begin{bmatrix} 1 \\ x_u \end{bmatrix} = 0$$

$\uparrow z_u$

Recasting into weighted least squares:

$$g = \sum_{u \in P} \left(\frac{\rho'(r_u)}{r_u} \right) r_u z_u \qquad \Big\downarrow \ r_u \neq 0$$

$$= \sum_{u \in P} w_u r_u z_u = 0$$

- Why? WLS has closed form solution

$$\hat{\alpha} = \bar{y} - \hat{\beta}(\bar{x} - c), \quad \hat{\beta} = \frac{\sum w_u (x_u - \bar{x})(y_u - \bar{y})}{\sum w_u (x_u - \bar{x})^2}$$

Q: How do we find $\hat{\alpha}, \hat{\beta}$?

A: Choose initial values → update residuals → keep going!

Also:

1. Init: $i \leftarrow 0$, $\hat{\theta}_0 = (\hat{\alpha}_0, \hat{\beta}_0)$
2. Loop:
   a) Get residuals & weights $\forall u \in P$

$$r_u = y_u - z_u' \hat{\theta}_i \rightarrow w_u = \frac{\rho'(r_u)}{r_u}$$

   b) Solve WLS problem via closed form expr. for $\hat{\alpha}$ & $\hat{\beta}$

$$\sum_{u \in P} w_u r_u z_u = 0$$

   c) Update param.

$$\hat{\theta}_{i+1} = \hat{\theta} \text{ (from b)}$$

   d) Check convergence b/w $\hat{\theta}_{i+1}$ & $\hat{\theta}_i$

Works for any θ vector in linear regression model: $y_u = z_u' \theta + r_u$

Newton - Raphson

Goal: Find $\theta \in \mathbb{R}^n$ s.t. $\Psi(\theta;\ P) = 0 \Rightarrow \Psi(\theta;\ P) = \begin{bmatrix} \Psi_1 \\ \vdots \\ \Psi_k \end{bmatrix} \Rightarrow k$ equations

Q: What's derivative of Ψ?

$$\Psi'(\theta; \rho) = \frac{\partial \Psi(\theta; \rho)}{\partial \theta} = \begin{bmatrix} \frac{\partial \Psi_1}{\partial \theta_1} & \cdots & \frac{\partial \Psi_1}{\partial \theta_k} \\ \vdots & & \vdots \\ \frac{\partial \Psi_n}{\partial \theta_1} & & \frac{\partial \Psi_n}{\partial \theta_k} \end{bmatrix} \rightarrow \text{Jacobian of } \Psi$$

If we have a guess $\hat{\theta}_i$, then near $\hat{\theta}_i$, we can say

$$\Psi(\theta; \rho) \approx \Psi(\hat{\theta}_i; \rho) + \Psi'(\hat{\theta}_i; \rho) \times (\theta - \hat{\theta}_i)$$

Setting this to 0

$$\Psi(\hat{\theta}_i; \rho) + \Psi'(\hat{\theta}_i; \rho)(\theta - \hat{\theta}_i) = 0$$

$$\therefore \hat{\theta}_{i+1} = \hat{\theta}_i - [\Psi'(\hat{\theta}_i; \rho)]^{-1} \Psi(\hat{\theta}_i; \rho)$$

Algorithm:

1. Initialize: $i \leftarrow 0$, $\hat{\theta}_0$
2. Loop:
   a) Update iterate:
   $$\hat{\theta}_{i+1} = \hat{\theta}_i - [\Psi'(\hat{\theta}_i; \rho)]^{-1} \Psi(\hat{\theta}_i; \rho)$$
   b) Check convergence or $i \leftarrow i+1$
3. Return $\hat{\theta}_i$

Note: If $\Psi = \nabla p$ (in grad descent), we are practically doing the same thing but descent controlled by Ψ'

SAMPLES

Idea: we cannot expect to have all population data to calculate attributes
↳ Use samples! Attributes of sample estimate pop. attributes

Samples introduce:

① Sample error:

$$\text{Sample error} = a(P) - a(S)$$

(P: pop., S: sample)

Depends on <u>attribute</u> & <u>sample</u>

We want to reduce this

② Consistency (Fischer)

If sample was pop ⟹ should yield some result

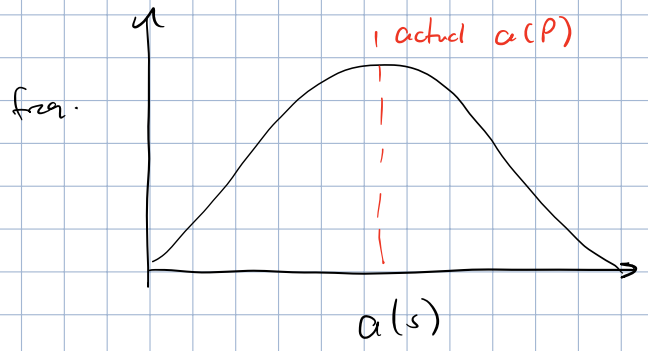
↳ i.e. sample error → 0 as n ↑

All possible samples

Idea: error depends on sample, so look @ all possible samples

# of samples: $\binom{N}{n}$ — N: total pop. size; n: desired sample size

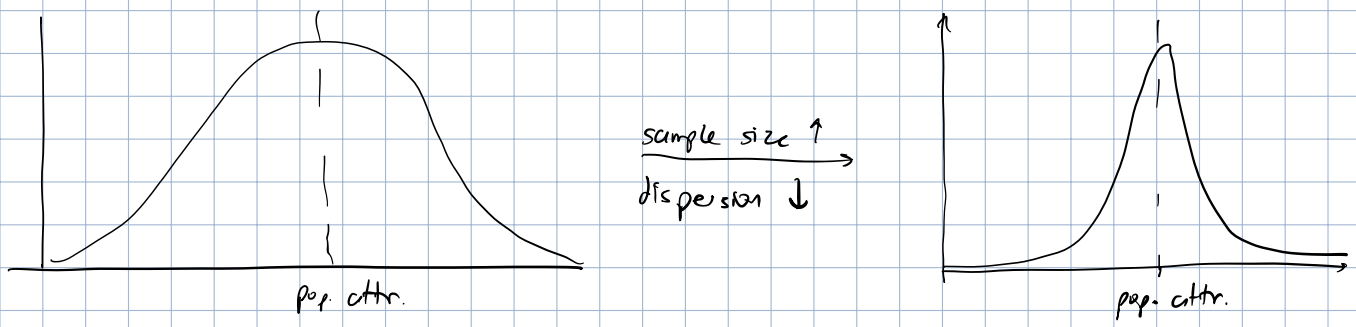
Calculate attribute on all possible samples



Taking sample error ∀ samples ⇒ avg. sample error ~ 0

Consistency & Effect of Sample Size

Sample error is dependent on sample size



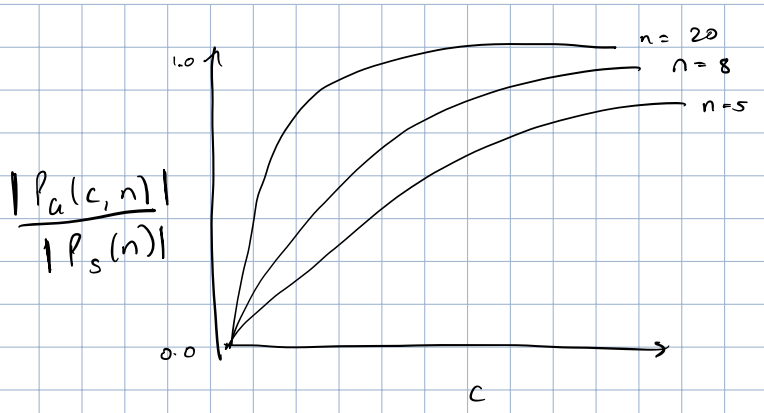
Q: How do we quantify this concentration?

Count # of samples s s.t. $\|a(s) - a(P)\|_1 < c$

∴ Let $P_S(n) = \{ S : S \subset P \wedge |S| = n \}$

$P_a(c, n) = \{ S : S \subset P_S(n) \wedge \|a(s) - a(P)\|_1 < c \}$

Then interesting graph:



As $c \uparrow \rightarrow P_a(c,n) \rightarrow 1$ $\forall$ sample size

But if $n \uparrow \rightarrow P_a(c,n) \rightarrow 1$ faster

Conclusion: larger samples cluster around pop. attribute

Not necessary that all attributes behave like this → could have skew, variability or consistent bias

However, almost all approach pop. attr. as sample size ↑

Comparison across attributes

We use relative absolute sample error to compare diff. attributes

Defn: For any $c > 0$, let

$$P_a^*(c,n) = \left\{ s : s \subset P_s(n) \wedge \frac{\| a(s) - a(P) \|_1}{\| a(P) \|_1} < c \right\}$$

Define proportion $\forall c > 0$, $n \leq N$

consistency w.r.t. same pop. attribute

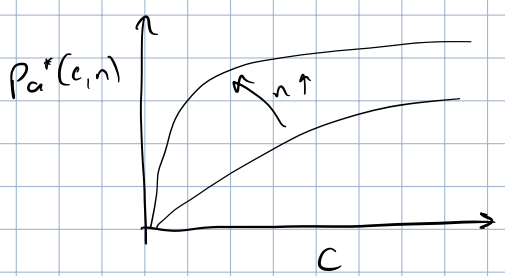
$$\left\{ P_a^*(c,n) = \frac{|P_a^*(c,n)|}{|P_s(n)|} \right.$$

→ # of samples of size n w/ relative absolute sample error $< c$

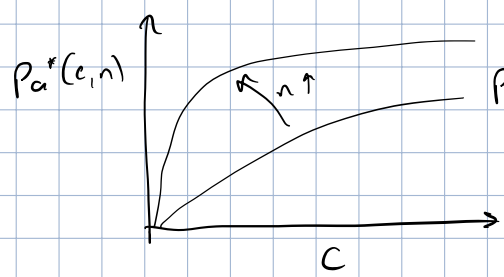
↳ # of samples of size

For cross-attribute comparisons, we evaluate how samples track pop. value

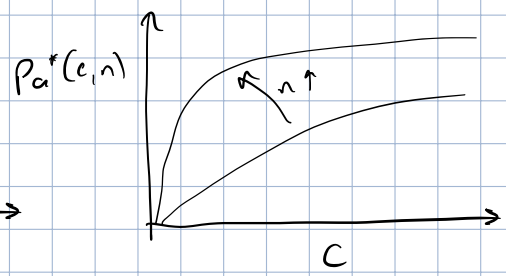
Attr. 1



Attr. 2



Attr. 3



The more left & above attr. α curves are compared to β, α outperforms β in consistency

These results may change based on population

Selecting Samples

Objective: select samples s.t. sample error is minimized

To help us, we use a sampling distribution. Properties are

① Exact ⇔ all possible samples available

② Approximate ⇔ subset of all possible samples available

③ In expectation ⇔ randomly choose 1 sample } - most realistic

② is often really good!

↳ Frequency histogram is scaled down but proportional density is identical!

Quantifying Sample Error

Recall that avg. sample error is:

$$\frac{1}{M} \sum_{s \in P_s} [a(s) - a(P)]$$

M: # of samples in P_s

P_s: population of all possible samples

We can quantify conc. of sample errors if we select a sample S via:

① Sampling bias

Expected sample error from repeated sampling of S from P_s

$$\text{Sampling bias} = E[a(s) - a(P)] = E[a(s)] - a(P)$$

$$= \sum_{s \in P_s} [a(s) - a(P)]\, p(s)$$

Result: If $p(s) = 1/M$, bias = avg. sample error of $a(P)$

Result: If bias = 0 ⇒ $a(s)$ is <u>unbiased estimator</u> of $a(P)$

② Sampling variance

Dispersion of sample errors

$$\text{Var}[a(s)] = E[(a(s) - E[a(s)])^2]$$

$$= \sum_{s \in P} [a(s) - E[a(s)]]^2 p(s)$$

Sampling standard deviation: $SD[a(s)] = \sqrt{\text{Var}[a(s)]}$

① Sampling mean squared error

To quantify distance b/w $a(S)$ & $a(P)$:

$$MSE[a(s)] = \underbrace{Var[a(s)] + Bias[a(s)]^2}_{\text{Should be as small as possible!}}$$

We can write all of this in terms of random variable where:

$$P(A = a) = \sum_{s \in P(s)} p(s) I[a(s) = a]$$

Sampling Mechanisms

We want to create a mechanism to sample based on units of a pop., not selecting a sample from a population of samples

Sampling mechanism defined by:

① $P(u)$: selecting unit u

② $P(u \mid k, s_{k-1})$: selecting u on k^{th} draw given we've observed sequence $s_{k-1} = (u_{i_1}, \ldots u_{i_{k-1}})$, u is k^{th} one in sample

To determine $p(s)$ (prob. of selecting sample s): sum $P(s_n)$ over all perms. of s_n

Mechanism 1: Simple Random Sampling Wout Replacement

① $Pr(u) = 1/N$

② $P(u \mid k, s_{k-1}) = \frac{1}{N-k+1} = \frac{1}{N-(k-1)}$

Chose $k-1$ units already, so $N-(k-1)$ units left

∴ Probability of sequence s_n:

$$Pr(s_n) = \frac{1}{N} \cdot \frac{1}{N-1} \cdot \frac{1}{N-2} \cdots \frac{1}{N-n+1}$$

Probability of selecting sample of size n

$$\frac{1}{\binom{N}{n}} = \frac{1}{M}$$

Mechanism 2: Simple Random Sampling w/ Replacement

① $P(u) = 1/N$

② $P(u \mid k, s_{k-1}) = \frac{1}{N}$

$\therefore$ Probability of S_n: $\Pr(S_n) = \left(\frac{1}{N}\right)^n$

$\therefore$ Probability of sample s: $\dfrac{\left(\frac{n!}{n_1! n_2! \dots n_N!}\right)}{N^n}$ $\Leftarrow$ n_k is # of duplicates of unit k

Mechanism 3: Simple Random Sampling - Weird Hybrid

Idea: do sampling w/ replacement but remove duplicates

↳ Sample size b/w 1 & n

Unit Inclusion Probabilities

Inclusion probability of unit u:

$$\pi_u = P(u \in S) = \sum_{s \in P_s} p(s) I[u \in s]$$

$D_u = \begin{cases} 1 & \text{if } u \in S \\ 0 & \text{o.w.} \end{cases}$

Thus: $P(D_u = 1) = \pi_u$, $P(D_u = 0) = 1 - \pi_u$

↳ $E[D_u] = 1 \cdot P(D_u = 1) + 0 \cdot P(D_u = 0)$
$= \pi_u$

↳ $\text{Var}[D_u] = E[D_u^2] - E[D_u]^2$

$= \left[1^2 \cdot P(D_u = 1) + 0^2 \cdot P(D_u = 0)\right] - \pi_u^2$

$= \pi_u - \pi_u^2$

Joint inclusion probability: both u & v are in sample

$\pi_{uv} = \Pr(u \in S \wedge v \in S)$

$= \sum_{s \in P_s} p(s) I[u \in s, v \in s]$

$\therefore \text{Cov}[D_u, D_v] = E[D_u D_v] - E[D_u] E[D_v]$

$= \pi_{uv} - \pi_u \pi_v$

Simple random sampling w/out replacement:

$$\pi_u = \frac{1 \cdot \binom{N-1}{n-1}}{\binom{N}{n}}, \quad \pi_{uv} = \frac{1 \cdot 1 \cdot \binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

Simple random sampling w replacement & hybrid

$\pi_u = 1 - P(u \notin S) = 1 - \left(\frac{N-1}{N}\right)^n$

$\pi_{uv} = 1 - P(u \notin S \vee v \notin S \vee u, v \notin S) = 1 - 2\left(\frac{N-1}{N}\right)^n + \left(\frac{N-2}{N}\right)^n$

Estimating Totals

Many attributes are either totals / function of totals

We can estimate the population total attr ($a(P) = \sum_{u \in P} y_u$) using Horvitz-Thompson estimator:

$$\hat{a(P)} = \sum_{u \in S} \frac{y_u}{\pi_u} = \sum_{u \in P} \frac{y_u}{\pi_u} \cdot D_u$$

Bias of estimator:

$$E\left[\tilde{a}_{HT}(S) - a(P)\right] = E\left[\tilde{a}_{HT}(S)\right] - a(P)$$

$$= E\left[\sum_{u \in P} \frac{y_u}{\pi_u} D_u\right] - a(P)$$

$$= \sum_{u \in P} \frac{y_u}{\pi_u} E[D_u] - a(P)$$

$$= \sum_{u \in P} \frac{y_u}{\pi_u} \cdot \pi_u - a(P)$$

$$= a(P) - a(P)$$

$$= 0$$

Variance:

Recall: $\mathrm{Var}\left[\sum a_i x_i\right] = \sum_i \sum_j a_i a_j \mathrm{Cov}(x_i, x_j)$

$\therefore \mathrm{Var}\left[\tilde{a}_{HT}(S)\right] = \mathrm{Var}\left[\sum_{u \in P} \frac{y_u}{\pi_u} D_u\right]$

$$= \sum_{u \in P} \sum_{v \in P} \frac{y_u}{\pi_u} \cdot \frac{y_v}{\pi_v} \cdot \mathrm{Cov}(D_u, D_v)$$

$$= \sum_{u \in P} \sum_{v \in P} \frac{y_u}{\pi_u} \cdot \frac{y_v}{\pi_v} \underbrace{(\pi_{uv} - \pi_u \pi_v)}_{\Delta_{uv}}$$

Yates-Grundy Formulation:

$$\text{Var}\left[\tilde{a}_{HT}(S)\right] = \frac{1}{2}\sum_{u\in P}\sum_{v\in P}\Delta_{uv}\left(\frac{y_u}{\pi_u}-\frac{y_v}{\pi_v}\right)^2$$

* Note: since bias = 0, MSE = variance for this estimator

We can actually estimate the variance of HT by using an HT estimate:

$$\text{Var}\left[\tilde{a}_{HT}(S)\right] = \sum_{u\in P}\sum_{v\in P}\Delta_{uv}\frac{y_u}{\pi_u}\frac{y_v}{\pi_v} = \sum_{(u,v)\in P(u,v)} q_{u,v}$$

$q_{u,v} = \Delta_{uv}\frac{y_u}{\pi_u}\cdot\frac{y_v}{\pi_v}$

Thus:

$$\widehat{\text{Var}\left[\tilde{a}_{HT}(S)\right]} = \sum_{(u,v)\in S_{u,v}}\frac{q_{u,v}}{\pi_{uv}} = \sum_{u\in S}\sum_{v\in S}\left(\frac{\pi_{uv}-\pi_u\pi_v}{\pi_{uv}}\right)\frac{y_u}{\pi_u}\frac{y_v}{\pi_v}$$

$S_{u,v}$ → sample of all possible pairs

Note:

$$\widehat{SE}\left[\tilde{a}_{HT}(S)\right] = \widehat{SD}\left[\tilde{a}_{HT}(S)\right] = \sqrt{\widehat{\text{Var}}\left[\tilde{a}_{HT}(S)\right]}$$

Why do we care? Via HT, we get:

① Estimate of f(pop. total)
② Estimate of variance of f(pop. total)

} unbiased!

Sampling Design

Defn:

$$(P_S, p(S))$$

P_S: Pop. of samples; $p(S)$: Probability of sample selected

We can choose samples & $p(S)$ s.t. MSE is low

From HT:

$$\text{MSE}\left[\tilde{a}_{HT}(S)\right] = \text{Var}\left[\tilde{a}_{HT}(S)\right] = -\frac{1}{2}\sum_{u\in P}\sum_{v\in P}\Delta_{uv}\left(\frac{y_u}{\pi_u}-\frac{y_v}{\pi_v}\right)^2$$

Choose sampling design to make this close to 0

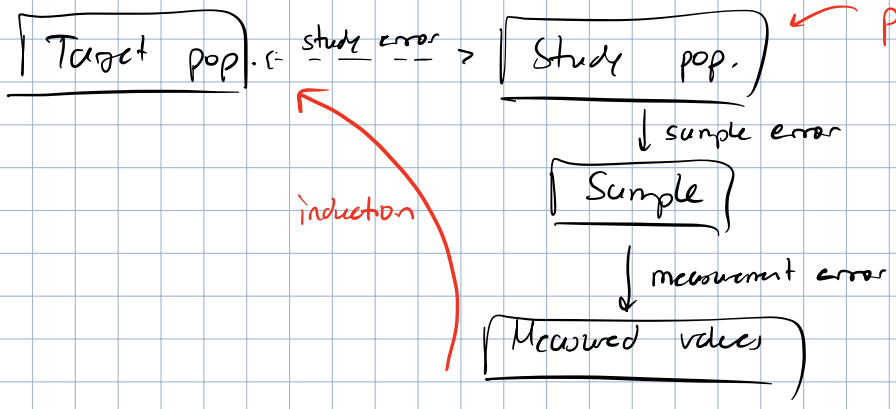
Ways to make MSE → 0:

① $\pi_u \propto y_u$
② $\pi_u \propto x_u$ if x_u correlates w/ y
③ $y_u \approx y_v \Rightarrow \pi_u \approx \pi_v$

Inductive Inference

If we have probabilistic sampling, we know relative freq. of certain attr. being selected → like an insurance policy

Sources of Error



Study error: $a(P_{study}) - a(P_{target})$ ← cannot be controlled by prob sampling

Total error from sample S:

$$a(S) - a(P_{target}) = [a(S) - a(P_{study})] + [a(P_{study}) - a(P_{target})]$$

Sources of measurement error:

① Measuring device bias/variability
② Operator error
③ Method error (eg. measuring length horizontal vs vertically)

Comparing Sub-populations

Goal: compare attributes of 2 populations, $a(P_1)$ and $a(P_2)$

First, let's define a measure of diff. b/w 2 attr.:

A: Measure of location attr.: $a(P_1) - a(P_2)$

B: Measure of spread IL: $a(P_1) / a(P_2)$

Q: If we were to figure out the diff, how do we know if it's bigger than the diff. observed if $P_1 \approx P_2$?

↳ Q: What does it mean for $P_1 \approx P_2$?

$P_1 = \{x_1, \ldots, x_n\}$
$P_2 = \{y_1, \ldots, y_n\}$

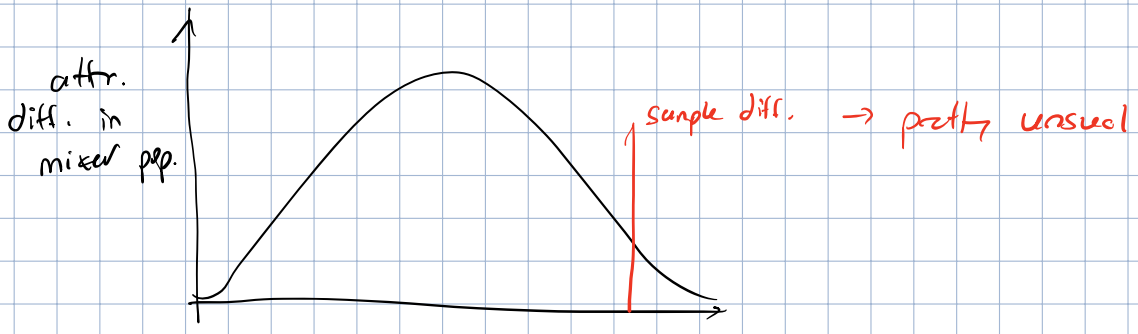
→ random mix into new pop. & draw 2 new pop. →

$P_1^* = \{y_1, x_2, \ldots, y_n\}$
$P_2^* = \{x_1, y_2, \ldots, x_n\}$

If attr. don't change, then $P_1 \approx P_2$

A: To know if diff. is unusual → hypothesis test for quantification

Intuition: draw many samples from mixed pop. & put attr. diffs. in histogram. If sample diff. is unusual, then we won't see very many instances of sample diff. in histogram



Anatomy of a Significance Test

Goal: quantify diff. in $a(P_1)$ vs. $a(P_2)$ in a randomly mixed pop.

Steps:

① Null hypothesis (H_0):

Hypothesize P_1 & P_2 from same pop. → $a(P_1) \sim a(P_2)$

H_A: alternative hypothesis ($\overline{H_0}$)

* Note: we do not state $H_0: a(P_1) = a(P_2)$, we say H_0 is 2 pop. drawn randomly from same pop. *

② Discrepancy measure

Quantifies inconsistency b/w observed data & H_0

↳ Large values of measure → evidence against H_0

Notation: $D(P_1, P_2)$

Ex:// 1. Average:

$$D(P_1, P_2) = |\bar{y}_1 - \bar{y}_2|$$

2. SD:

$$D(P_1, P_2) = \left| \frac{SD(P_1)}{SD(P_2)} - 1 \right|$$

3. Pop. avg of P_1 is greater than pop. avg. of P_2

$$D(P_1, P_2) = \bar{y}_2 - \bar{y}_1$$

④ Calculate observed discrepancy

d_{obs}

* Note: this is only looking @ 1 attr of diff. *

⑤ Compare observed discrepancy against sampling distr. of discrepancy from randomly mixing pop.

⑤ Calculate p-val

$$p\text{-val} = Pr(D \geq d_{obs} \mid H_0 \text{ is true})$$

If p-val small, 2 options:

A: H_0 is true & we observed very unusual data
B: H_0 is false

} Choose to believe B

To calculate p-val exactly: take all permutations of pop & derive sampling distr.

↳ Approximation: just take M random sub-pops,

$$\hat{p\text{-val}} = \frac{1}{M} \sum_{i=1}^{M} I\left[D(p_{1,i}^*, p_{2,i}^*) \geq d_{obs}\right]$$

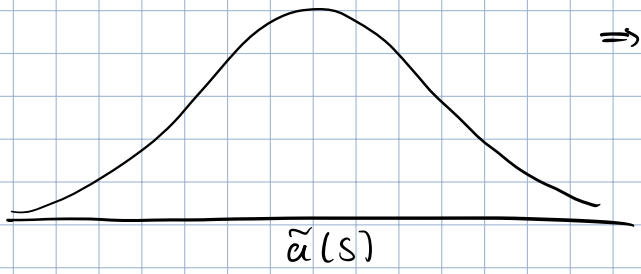
Count # of random mixed sub pops. with discrepancy $\geq d_{obs}$

Errors in sig. tests:

① Type I: reject H_0 but actually H_0 true

② Type II: accept H_0 but actually H_0 false

Revisiting Sampling Distributions



⇒

① Construct all possible samples of particular size

② Took a(s) per sample

③ Distr.

To find 95% of all possible $\tilde{a}(s)$ ⇒ take 2.5th & 97.5th quantiles of sampling distr.

(all possible: Too much!)

Q: How can we recreate sampling distr. w/out finding all possible samples

A: 3 methods:

① Take large # of samples & approx.

② Approx to normal distr. (why? CLT!)

↳ Useful if attr. is a function of total, O.W. not guaranteed.

③ Approx via resampling method

Random vs. Observed Intervals

If pop. attr. of interest is $a(P) = \mu = \frac{1}{N} \sum_{u \in P} y_u$, then estimator from sample is:

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \sigma^2 = \frac{1}{N} \sum_{u \in P} (y_u - \mu)^2$$

(CLT)

Standardized:

$$Z = \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

* Important: if pop. is finite, include finite pop. corr.

$$\text{Var}[\bar{Y}] = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$ (assuming σ^2 is known from pop.)

↳ If $N >> n$, $\frac{N-n}{n-1} \to 1$

↳ For completeness:

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right), \quad Z = \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} \sim N(0,1)$$

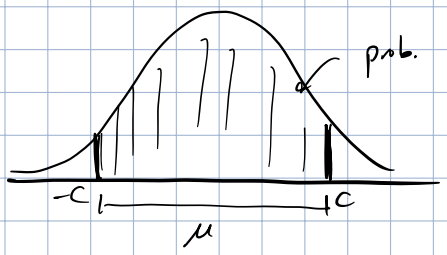
Random interval:

Want to find interval that contains μ w/ prob. $1-p$ (coverage prob.)

$$\left[\bar{Y} - c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}, \quad \bar{Y} + c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right]$$

Random b/c defined by R.V. Can substitute $\bar{Y}$ w/ $\bar{y}$

Derivation:



$$1-p = Pr\left(-c \le \frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} \le c\right)$$

$$= Pr\left(\bar{Y} - c\ldots \le \mu \le \bar{Y} + c\ldots\right)$$

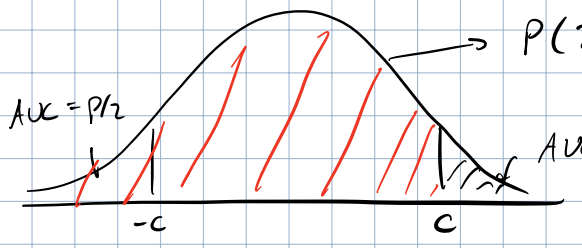
Meaning: μ is covered by interval $100(1-p)\%$ of time

How to determine c?

$$c = Q_z\left(1 - \frac{p}{2}\right)$$

↑ z distr.

Why?



$$P(Z \le c) = 1 - p/2$$

$$= Q_z(1 - p/2)$$

Observed interval

We usually only have 1 sample → switch $\bar{Y}$ to $\bar{y}$

$$\left[\bar{y} - c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}},\ \bar{y} + c \times \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right]$$

Known as confidence interval. Not random ⇒ either has or doesn't have μ

If normality assumption T ⇒ $100(1-p)\%$ of intervals will contain μ

Student t-based intervals

Assumption: know $SD[\bar{Y}]$ → unrealistic

We can estimate!

↳ In HT, we saw $SE[\hat{a}(S)] = \widehat{SD}[\hat{a}(S)]$

∴ Pivotal quantity: $\frac{a(S) - a(P)}{SE[\tilde{a}(S)]}$

If data is normally distr., then:

$$T = \frac{\bar{Y} - \mu}{SE[\bar{Y}]} = \frac{\bar{Y} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} \sim t_{(n-1)}$$

(n: sample size)

Interval via t-distr:

$$\left[\tilde{a}(S) - c \times \hat{SD}[\hat{a}(S)],\ \tilde{a}(S) - c \times \hat{SD}[\hat{a}(S)]\right] \Rightarrow$$

($\tilde{a}(S)$: random; $\hat{SD}[\hat{a}(S)]$: random)

Random center & random length!

How do we calculate c?

↳ Same idea, but use $Q_{T_{(n-1)}}(1 - \alpha/2)$ instead of Q_Z

For pop. avg:

Random interval:

$$R.I. = \bar{Y} \pm c \times \frac{\tilde{\sigma}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

($\tilde{\sigma}$ is R.V. from sampling distr.)

Observed:

$$C.I. = \bar{y} \pm c \times \frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \Rightarrow$$

($\hat{\sigma}$ is observed: $\hat{\sigma} = \sqrt{\frac{\sum(y_n - \bar{y})}{n}}$)

Pivotal quantities are important!

↳ For scale attr.: $\frac{\tilde{s}(S)}{s(P)} \sim \chi^2_{\text{form}}$

Resampling Intuition

Problem: assuming we know sampling distr. of discrepancy / pivot

Q: How can we figure out sampling distr. w/out taking all possible samples?

A: ① Take M samples & take attr. → find distr.

↳ Problem: requires repeated sampling. We usually only have 1 sample.

② Take B samples from original sample $\underline{S}$ rather than pop. !

Using S as approx. of study pop. !

Need to sample w/ replacement since S is small

Each sample from S is <u>bootstrap sample</u>

Take attr. from each sample & find distr.

Bootstrap Method

Bootstrap estimate of distr. = bootstrap samples → take attr. → find distr of attr.

$$P_{target} \longleftrightarrow P_{study} \longleftrightarrow P^* \approx S$$

$$P_{study} \downarrow S \qquad P^* \approx S \downarrow S^*$$

V. powerful method of estimating distr. w/ only 1 sample

Depends if S is good approx of P → can only be as good as a(S)

Bootstrap sample error: $a(S^*) - a(S)$

↑ bootstrap sample

Bootstrap standard deviation:

$$SE = \hat{SD}_* [\tilde{a}(S)] = \sqrt{\frac{\sum_{b=1}^{B} [a(S_b^*) - \bar{a}^*]^2}{B-1}} \Rightarrow \bar{a}^* = \frac{1}{B}\sum_{b=1}^{B} a(S_b^*)$$

↳ For pop. avg:

$$\hat{SD}_* [\bar{Y}] = \sqrt{\frac{\sum_{b=1}^{B} (\bar{y}_b^* - \bar{y}^*)^2}{B-1}} \Rightarrow \bar{y}_b^*: \text{ avg of bootstrap sample } b$$

$$\bar{y}^*: \frac{1}{B}\sum_{b=1}^{B} \bar{y}_b^*$$

Compared to the original method of estimating $SD[\bar{Y}]$, bootstrap slightly overestimates

Pros of bootstrap distribution:

① Estimating sampling bias

$\hat{\text{sampling bias}}$ = avg. bootstrap sample error = $\bar{a}^* - a(S)$

If we know estimator is biased, then we can correct by defining new cstr.

$$a^*(s) = a(s) - \text{bias}$$

Using estimate of bias:

$$a^*(s) = a(s) - \bar{a}^* + a(s)$$
$$= 2a(s) - \bar{a}^*$$

② C.I. & hypothesis testing

Bootstrap Confidence Intervals

We will use bootstrap distr. as proxy for sampling ⇒ approx. C.I.

Naive normal theory intervals:

C.I. now defined as $\bar{y} \pm c * \underbrace{\hat{SD}[\bar{Y}]}_{\text{bootstrap estimate}}$

(still from normal dist.)

Percentile method

① Generate B bootstrap samples from sample S

② For each sample, calculate $a_b = a(S_b^*)$ ↦ cstr.

③ Take quantiles using $a_1, \ldots, a_B$

④ Code C.I.

This is equivariant to any transformation of attributes

Pro: simple & transform equivariant

PREDICTION

Accuracy of Prediction

Obj: determine if model $\hat{y} = \mu(\hat{x})$ ($y = \mu(x) + \text{error}$) is accurate

Average prediction squared error (APSE):

$$\frac{1}{N} \sum_{\text{all } i} (y_i - \hat{\mu}(x_i))^2$$

} Quantifies distance b/w actual & predicted val.

↳ Proportional to residual mean squared error ($\sum_{i=1}^{N} r_i^2$)

↳ Lower APSE → better prediction

We often notice that making model more complex leads to lower APSE but becomes overtuned

We should use APSE on data not used to train the model:

$$APSE(P, \hat{\mu}_S) = \frac{1}{N} \sum_{u \in P} (y_u - \hat{\mu}_S(x_u))^2$$

$\hat{\mu}_S$: Train model on sample S

P: data to calculate APSE on; S: data for model calc.

$$= \left(\frac{n}{N}\right) APSE(S, \hat{\mu}_S) + \left(\frac{N-n}{N}\right) APSE(T, \hat{\mu}_S)$$

$APSE(S, \hat{\mu}_S)$: APSE from sample

$APSE(T, \hat{\mu}_S)$: APSE from data not use in training

Usually we use:

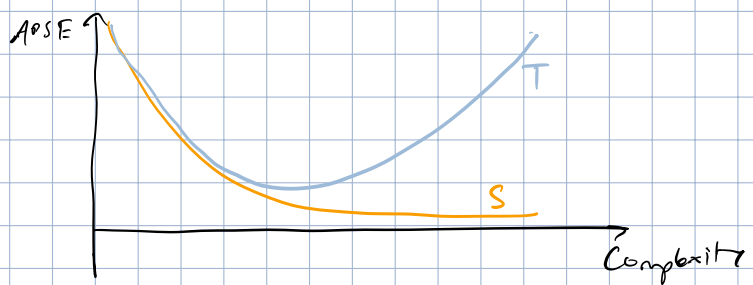
$$APSE(T, \hat{\mu}_S) = \frac{1}{N-n} \sum_{u \in T} (y_u - \hat{\mu}_S(x_u))^2$$

If $n \ll N \rightarrow APSE(T, \hat{\mu}_S) \approx APSE(P, \hat{\mu}_S)$

T: data not used in training

Usually $APSE(T, \hat{\mu}_S) > APSE(S, \hat{\mu}_S)$

Overfitting: make model too complex → not great @ predicting outside of sample



APSE highly depends on quality of sample we choose to use for training

Predictions over Multiple Samples

Problem w/ naive APSE: error depends on initial sample picked for fitting model

Soln:

① Take multiple samples & fit model $\hat{\mu}_{S_j}(\vec{x})$

② Take $APSE(P, \hat{\mu}_{S_j})\ \forall j$ → sampling distr. of APSE

③ Take avg. APSE

$$APSE(P, \tilde{\mu}) = \frac{1}{M} \sum_{j=1}^{M} APSE(P, \hat{\mu}_{S_j})$$

We can decompose $APSE(P, \bar{\mu})$ into 3 different parts:

↳ Preamble: $\tau(\vec{x})$ is the relationship b/w y & x observed in pop.

Since x can have duplicate vals. w/ diff. y-vals, we partition pop. into:

$$P = \bigcup_{k=1}^{K} A_k \quad \curvearrowleft \; A_k \text{ is cluster of units w/ same } x \text{ value}$$

For each cluster:

$$T(x_k) = \underbrace{\frac{1}{n_k}}_{\substack{\# \text{ of units} \\ \text{in cluster}}} \sum_{u \in A_k} y_u$$

↳ Decomposition:

Note: $\bar{\mu}(x) = \frac{1}{M} \sum_{j=1}^{M} \hat{\mu}_{S_j}(x) \Rightarrow$ Avg. over M estimated predictor func. across samples

$$APSE(P, \bar{\mu}) = \frac{1}{M} \sum_{j=1}^{M} \frac{1}{N} \sum_{u \in P} (y_u - \hat{\mu}_{S_j}(x_u))^2$$

$$= \frac{1}{M} \sum_{j=1}^{M} \frac{1}{N} \sum_{u \in P} (y_u - \tau(x_u))^2 + \underbrace{\frac{1}{M} \sum_{j=1}^{M} \frac{1}{N} \sum_{u \in P} (\hat{\mu}_{S_j}(x_u) - \tau(x_u))^2}_{\text{MSE-like!}}$$

$$= \text{"} + \underbrace{\frac{1}{M} \sum_{j=1}^{M} \frac{1}{N} \sum_{u \in P} (\hat{\mu}_{S_j}(x_u) - \bar{\mu}(x_u))^2}_{\text{Var}[\bar{\mu}]} + \underbrace{\frac{1}{N} \sum_{u \in P} (\bar{\mu}(x_u) - \tau(x_u))^2}_{\text{Bias}^2[\bar{\mu}]}$$

$\text{Ave}_x(\text{Var}[Y|x])$

⇓

ave. of conditional variance of y given x

Indep of $\mu(x)$

Note that:

$$APSE(P, \bar{\mu}) = \left(\frac{n}{N}\right) \left(APSE(P, \bar{\mu}) \text{ on samples used by } \mu\right) + \underbrace{\left(\frac{N-n}{N}\right) \left(APSE(P, \bar{\mu}) \text{ on samples not used by } \mu\right)}$$

Dominates if $n \ll N$, else more for since out-of-sample perf.