# intro

- foreshadowing/context: under all prob computations are *sample spaces*
- rarely work with sample spaces directly, unless they're simple (heads/tails)
  - (so, we start here)

# what is prob?

- objective prob: long run freq of occurrence (eg heads in coin flip)
  - often called frequentist/classical methods.
  - used more often in undergrad CMU
- subjective prob: a possibly informed belief in rate of occurrence of event
  - can called bayesian

# set notation

- $A \supset B, A \subset B, A \cup B, A \cap B, \overline{A}$ aka $A^C$
- let the set of all experimental outcomes $\Omega = A \cup \overline{A}$
- $A \cap B = \emptyset \Longrightarrow A$ and $B$ are mutually exclusive aka disjoin
- distributive/associative laws
- de morgan's ($\overline{A \cup B} = \overline{A} \cap \overline{B}$, etc)

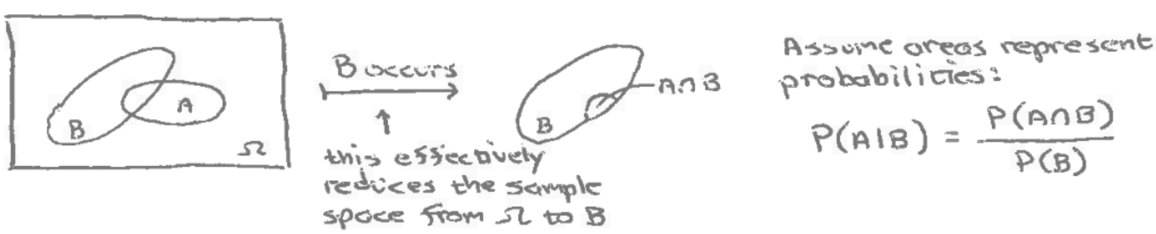
# what are experiments?

- make observations
- passive: just collect data
- active: control setting
- sample space ($\Omega$)
  - two coins tossed? $\Omega = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$
    - HH is simple event
    - TH is compound event ("at least one tail")
  - free throws until miss? $\Omega = \{\mathrm{M},\ \mathrm{HM},\ \mathrm{HHM},\ \ldots\}$
  - relative freqs of above? don't know! need more info

# probability

## sample space probabilities

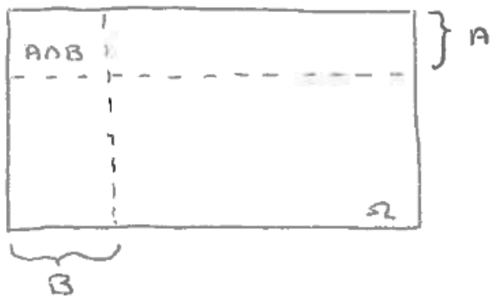
- conditional probability: $A$, when we know $B$




- $A$ and $B$ are independent
  - iff $P(A \mid B) = P(A)$
  - iff $P(B \mid A) = P(B)$
  - iff $P(A \cap B) = P(A)P(B)$
- if there are three events, $A$ and $B$ are *conditionally* independent

- if $P(A \cap B \mid C) = P(A \mid C)P(B \mid C)$

The following is an example of independence as rendered on a Venn diagram:



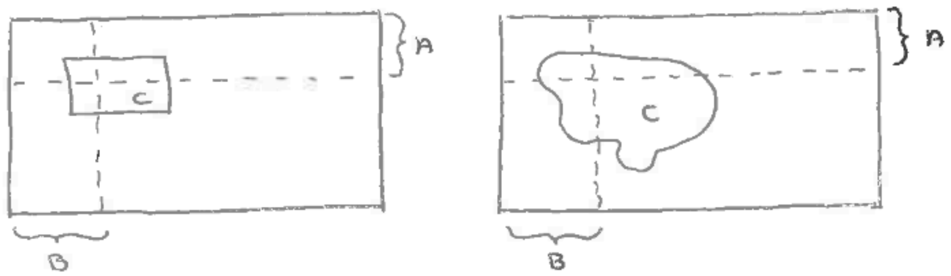

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(\Omega)} = P(A)$$

area ratios are equal

Let $A$, $B$, and $C$ be three separate events in $\Omega$, all of which have non-zero probability of occurring. $A$ and $B$ are *conditionally independent* if

$$P(A \cap B \mid C) = P(A|C)P(B|C)$$

The following shows conditional independence and dependence as rendered on a Venn diagram:




- multiplicative law:

$$\begin{aligned} P(A \cap B) &= P(A)P(B \mid A) = P(B)P(A|B) \\ &= 0 \text{ if } A, B \text{ disjoint} \\ &= P(A)P(B) \text{ if } A, B \text{ independent} \end{aligned}$$

We can generalize this result to $n$ events $\{A_1, \cdots, A_n\}$:

$$P(A_1 \cap \cdots \cap A_n) = P(A_1 \mid A_2 \cap \cdots \cap A_n)\underbrace{P(A_2 \cap \cdots \cap A_n)}$$

$$\downarrow$$

$$P(A_2 \mid A_3 \cap \cdots \cap A_n)\underbrace{P(A_3 \cap \cdots \cap A_n)}$$

$$\downarrow$$

$$P(A_3 \mid A_4 \cap \cdots \cap A_n)\underbrace{P(A_4 \cap \cdots \cap A_n)}$$

$$\downarrow$$

etc.

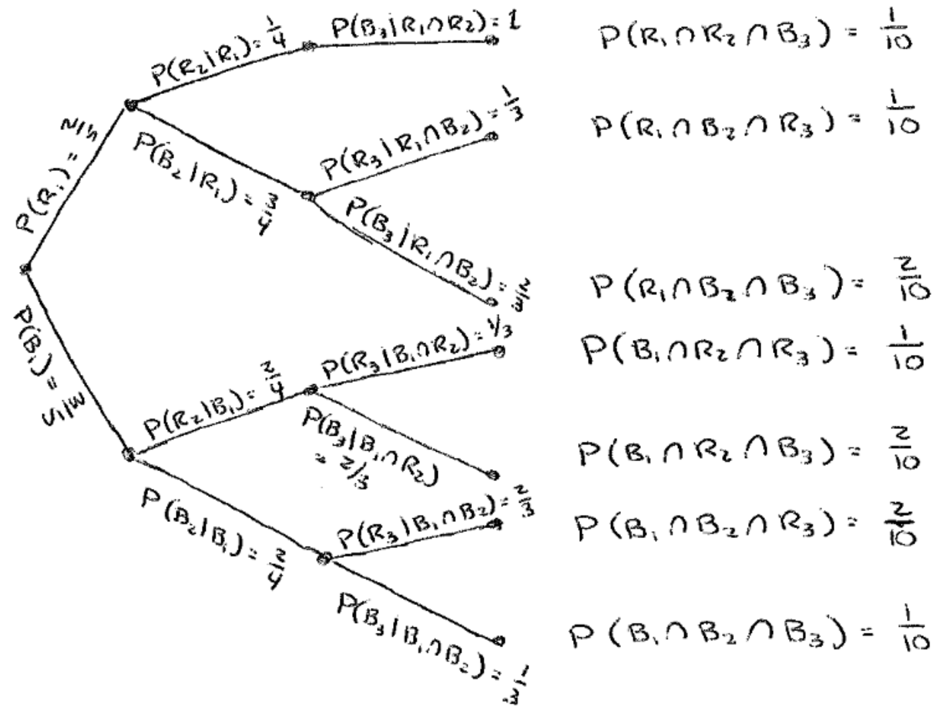
$$= P(A_n) \prod_{i=1}^{n-1} P(A_i \mid A_{i+1} \cap \cdots \cap A_n)$$

- additive law:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) \text{ if } A, B \text{ disjoint} \\ &= P(A) + P(B) - P(A)P(B) \text{ if } A, B \text{ independent} \end{aligned}$$

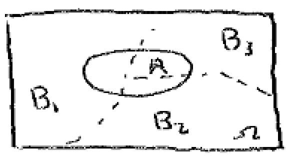
- decision trees: good for when probabilities change
  - eg picking colored balls without replacement, not like probability of heads of fair coin

→ **Example**: you pull three balls out from an urn which has two red and three black balls, without replacement. (a) What is the probability that the second ball drawn is red? (b) What is the probability that the second ball drawn is red, if the third ball drawn is black?




- for these next two, let $\{B_i\}$ be a partition of $\Omega$.
- law of total probability.

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$




  - "probability of $B_1$. then times the probability that $i$ landed in $A$, in $B_1$, etc etc"
  - helpful when given conditional probs but not $A$ itself

- Bayes' Rule:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)} = \frac{P(A|B_j)P(B_j)}{P(A)}$$

→ **Example: The Monty Hall Problem.** This problem is named for the long-time host of the game show *Let's Make a Deal.* The simple version goes as follows: you are shown three doors; behind two of the doors are goats and behind the other is a car. You choose a door (say Door #1). Monty Hall then opens, say, Door #3 to reveal a goat, and asks you if you want to switch to Door #2.

So: do you stick with Door #1 or switch to Door #2?

Assume Door #1 has been selected.

$O_i$ = "Monty opens Door $i$" $C_i$ = "car is behind Door $i$"

$P(C_i) = \frac{1}{3}$ for all $i$ ← car could be behind any door

$\Omega = \{ O_2 \cap C_1, \cancel{O_2 \cap C_2}, O_2 \cap C_3, O_3 \cap C_1, O_3 \cap C_2, \cancel{O_3 \cap C_3} \}$

Monty will <u>not</u> open the door the car is behind.

What is $P(C_2|O_3)$?

$$P(C_2|O_3) = \frac{P(O_3|C_2)P(C_2)}{P(O_3)} = \frac{P(O_3|C_2)P(C_2)}{P(O_3|C_2)P(C_2) + P(O_3|C_1)P(C_1)}$$

$P(O_3|C_2) = 1$ (no choice) $P(O_3|C_1) = \frac{1}{2}$

$$\Rightarrow P(C_2|O_3) = \frac{1 \cdot 1/3}{1 \cdot 1/3 + 1/2 \cdot 1/3} = \boxed{\frac{2}{3}}$$

⇒ change your pick to Door # 2!

## other probability paradigm??

- need better paradigm for continuous situations/nondiscrete outcomes: random variables, prob distributions
- random variable: like a function, maps events in Ω to, eg, real number line (the output. meters ran maybe)
  - but tbh just think of random variables as their outputs
  - can have diff random variables for same experiment: number of tails observed, number of heads, whether not heads observed
  - denoted with uppercase Latin eg $X$, $P(X = x)$
  - NO INHERENT NOTION OF PROBABILITY!
- functions of random variables are random variables! *statistics* are *functions of data* are *random variables*
- properties of discrete and continuous prob distributions

| | Discrete | Continuous |
|---|---|---|
| Name | pmf: probability mass function | pdf: probability density function |
| Symbol | $p_X(x)$ | $f_X(x)$ |
| Properties | $0 \le p_X(x) \le 1$ | $f_X(x) \ge 0$ |
| | $\sum_{\text{all } x} p_X(x) = 1$ | $\int_{\text{all } x} f_X(x) dx = 1$ |
| | $P(a \le X \le b) = \sum_{x \in [a,b]} p_X(x)$ | $P(a \le X \le b) = \int_a^b f_X(x) dx$ |
| | $P(a < X < b) = \sum_{x \in (a,b)} p_X(x)$ | $P(a < X < b) = \int_a^b f_X(x) dx$ |
| | $E[X] = \sum_{\text{all } x} x p_X(x)$ | $E[X] = \int_{\text{all } x} x f_X(x) dx$ |
| (law of unconscious statistician) | $E[g(x)] = \sum_{\text{all } x} g(x) p_X(x)$ | $E[g(x)] = \int_{\text{all } x} g(x) f_X(x) dx$ |

- expected value operator $E[X] = \mu_X$, mean of the distribution $X$ was sampled from
  - $E[cX] = cE[X]$
  - $E[c] = c$
  - $E[x + y] = E[x] + E[y]$
- variance operator $V[X] = \sigma^2$
  - $V[X] = E\left[(x - \mu)^2\right] = \text{(simplifies to) } E[X^2] - (E[X])^2$
  - apparently $V[x + y] = V[x] + V[y]$??? sep 18 class example 2
  - variance is not width, but the square of the width. think about units of $V[X]$ vs units of $E[X]$
- translation/scaling's effects on mean and variance: $(X \to X + b)$
  - $E[X + b] = E[X] + b$ (translation shifts mean)
  - $E[aX] = aE[X]$ (scaling shifts mean multiplicatively)
  - $V[X + b] = V[X]$ (translation doesn't effect width)
  - $V[aX] = a^2 V[X]$ (scaling widens exponentially. verify via shortcut formula)

- cumulative distribution function (cdf)
  - accumlated prob up to $x$, inclusive

| | Discrete | Continuous |
|---|---|---|
| Symbol | $F_X(x)$ | $F_X(x)$ |
| Definition | $F_X(x) = \sum_{y \le x} p_Y(y)$ | $F_X(x) = \int_{-\infty}^{x} f_Y(y)dy$ |
| Limiting Properties | for both, $F_X(-\infty) = 0$ and $F_X(\infty) = 1$ | |
| Reln to pmf/ pdf | $p_X(x)$ is magnitude of jump in $F_X(x)$ at coord $x$ | $f_X(x) = \frac{d}{dx} F_X(x)$ |
| Reln to Quantile $q$ | $X = \min\{x : F_X(x) \ge q\}$ | inverse cdf, $X = F_X^{-1}(q)$ |
| Reln to Prob. Over Range | It's complicated. | $P(a \le X \le b)$ $= P(a < X < b)$ $= F_X(b) - F_X(a)$ |

  - inverse cdf: given total prob $q$ to left of (and including) some $x_0$, what is $x_0$?

| | input | output |
|---|---|---|
| cdf | $x_0$ | $q$ |
| inverse cdf | $q$ | $x_0$ |

**families of distributions**

- up till now, practice problems have fixed $\theta$ in equations. we often have to find it. (use law of total prob)
- deriving from Law of Total Prob ($P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$) if we know
  - $p_{X|\theta}(x|\theta)$
  - with $\theta$ weights from $p_\Theta(\theta)$,

$$p_X(x) = \sum_\theta p_{X|\theta}(x|\theta) p_\Theta(\theta)$$

  - if $\theta$ is continuous, we may just use an integral
  - 

Now, what happens if $X$ is a continuous random variable?

Discrete $\Theta$ : $f_X(x) = \sum_{\text{all } \Theta} f_{X|\Theta}(x|\Theta)\, p_\Theta(\Theta)$

Continuous $\Theta$ : $f_X(x) = \int_{\text{all } \Theta} f_{X|\Theta}(x|\Theta)\, f_\Theta(\Theta)\, d\Theta$

- plotting such a pdf: $f_X(x) = \int_0^\infty \theta x^{\theta-1} \cdot e^{-\theta} d\theta$

```
x   <- seq(0.001,1,by=0.001)
f.X <- rep(NA,length(x))

f <- function(theta,x) {
  return(theta*x^(theta-1)*exp(-theta))
}
for ( ii in 1:length(x) ) {  # we loop over x's indices (1, 2, ..., length(x))
  f.X[ii] <- integrate(f,0,Inf,x=x[ii])$value
}

# Now let's plot!
library(ggplot2)
df <- data.frame(x,f.X)
ggplot(data=df,mapping=aes(x=x,y=f.X)) +
  geom_line(col="firebrick") +
  ylim(c(0,3)) +
  ylab(expression(f[X]*"(x)"))
```

### data sampling code

- inverse-transform sampling
  1. sample a $q \in (0, 1)$. (e.g. `runif()`, random uniform)
  2. plug $q$ into $x = F_X^{-1}(q)$, record $x$
  3. repeat $n$ times for a sample size of $n$
- rejection sampling
  1. choose finite domain $[a, b]$ for $f_X(x)$ that's good enough
  2. let $\max(f_X(x)) = m$
  3. repeat until $n$ samples recorded:
     1. randomly sample $x' \in [a, b]$ and $y' \in [0, m]$.
     2. if $y' \leq f_X(x')$, keep the data point. otherwise, reject it and continue
- see `media/data_sampling.Rmd`

### statistics

- iid: independant and identically distributed
- reiterating: these are just functions of observed data
  - $Y = X_1 + \cos(X_2) - \frac{X_{37}}{\pi}$ is a statistic, but not informative
  - $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_1$ is as well, and is informative of $\mu$
  - $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_1 - \overline{X}\right)^2$ is informative of $\sigma^2$
- statistics are random variables!!!
  - drawn from sampling distributions (which are just pmfs/pdfs)file:
- sample mean, stddev, range, median
- $E\left[\overline{X}\right] = E[X]$, $V\left[\overline{X}\right] = \frac{V[X]}{n}$ hold for all distributions
- standard error is the width of the distribution; the standard deviation of a statistic; $\sqrt{V[Y]}$, where $Y$ is a random variable/sampling distribution/pmf or pdf for a statistic
- expected value of the sample variance $S^2$ is $\sigma^2$ after LOTS of math
  - this is Wednesday: Statistics and Sampling Distributions class example 1
  - why is $\frac{1}{n-1}$ there? create an unbiased example of the population estimate??

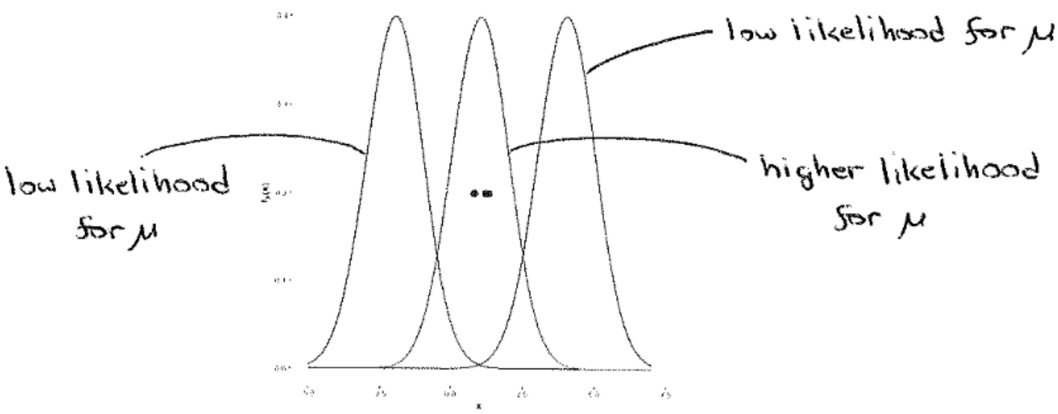
**likelihood function**

- the likelihood function quantifies how likely a $\theta$ is given our data
- let it be defined:

$$\mathcal{L}\left(\theta \mid \vec{X}\right) = \prod_{i=1}^{n} f_X(x_i \mid \theta)$$

  for continuous data, simply using $p_X(x_i \mid \theta)$ for discrete

- notice that




- why? in inference, we want to estimate $\theta$ given data! so we maximize $\mathcal{L}$ for $\theta$
- to make math easier, we use the log-likelihood $\ell\left(\theta \mid \vec{X}\right) = \log \mathcal{L}\left(\theta \mid \vec{X}\right)$

**bias and variance**

- bias: $B\left[\hat{\theta}\right] = E\left[\hat{\theta} - \theta\right] = E\left[\hat{\theta}\right] - \theta$
  - estimator biased: $B\left[\hat{\theta}\right] = 0$, unbiased $B\left[\hat{\theta}\right] \neq 0$
- variance: $V\left[\hat{\theta}\right]$ (recall defn above)
- mean-squared error (MSE): $\mathrm{MSE}\left[\hat{\theta}\right] = E\left[\left(\hat{\theta} - \theta\right)^2\right] \underset{\text{simpl to}}{=} V\left[\hat{\theta}\right] + \left(B\left[\hat{\theta}\right]\right)^2$
  - select "best" estimator by taking the one with lowest MSE

**Maximum Likelihood Estimation**

- procedure (it's literally just AP Calc maximization):
  1. find $\mathcal{L}\left(\theta \mid \vec{X}\right)$
  2. find $\ell\left(\theta \mid \vec{X}\right)$
  3. compute $\ell'\left(\theta \mid \vec{X}\right)$, partial or normal with respect to $\theta$
  4. solve the above equal to 0 for $\theta$, now called $\hat{\theta}_{\mathrm{MLE}}$ (also replace $x_i$ with $X_i$)
- this does not work with domain-specifying parameters (e.g. $f_X(x) = \frac{1}{\theta}$ for $x \in [0, \theta]$)
- property of MLEs: invariance property
  - if $\theta' = g(\theta)$, then $\hat{\theta}'_{\mathrm{MLE}} = g\left(\hat{\theta}_{\mathrm{MLE}}\right)$
  - e.g., .0 what

**more abt confidence intervals**

$P\left(\hat{\theta}_L \leq \theta \leq \hat{\theta}_R\right)$ where the theta HATS are the r.v.s!

point estimates dont need sampling distributions! point estimates do!

when doing the math, finding lower and upper bound should be the same thing except having a swapped $q$

**erm.**
we skipped a bunch of notetaking. it's exam 3 time.

**review of point estimation**

- moment gen fn $M_X(t) = E[e^{tX}]$
  - if $Y = b + \sum_{i=1}^{n} a_i X_i$, then $M_Y(t) = e^{bt} \sum_{i=1}^{n} M_{X_i}(a_i t)$
  - match mgfs to match distributions
  - used for CIs and HTs
    - sum of $n$ indep normal r.v.s is a normal r.v.
    - sample mean of $n$ iid normal r.v.s is a normal r.v.
    - standardized normal r.v. ($\frac{x-\mu}{\sigma}$) is a standard normal r.v.
    - sum of $n$ squared standard normal r.v.s is $\chi^2$ distr for $n$ deg of freedom
- general transformations of single random variable
  - given $f_X(x)$ and $U = g(X)$, what is $f_U(u)$?
  - why?
  - sqre of one std normal r.v. is $\chi^2$ for 1 d.o.f.
  - in part, allows derivation of $t$ distr.
- distribution of $T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1} \xrightarrow{d} Z \sim N(0,1)$ as $n \to \infty$
- performance of point estimators
  - bias: $B\left[\hat{\theta}\right] = E\left[\hat{\theta} - \theta\right] = E\left[\hat{\theta} - \theta\right]$ (if bias tends to 0, asymptotically biased.)
  - variance: $V\left[\hat{\theta}\right]$
  - mean-squared error: $\mathrm{MSE}\left[\hat{\theta}\right] = \left(B\left[\hat{\theta}\right]\right)^2 + V\left[\hat{\theta}\right]$
  - consistency: tldr, whether $\mathrm{MSE}\left[\hat{\theta}\right] \to 0$ as $n \to \infty$
- how can we get a good point estimator? Maximum Likelihood Estimation!
  - asymptotically unbiased
  - invariance property: $\widehat{g(\theta)}_{\mathrm{MLE}} = g\left(\hat{\theta}_{\mathrm{MLE}}\right)$
  - maximuze $l$ (log-likelihood function)

**Cramer-Rao Lower Bound (CRLB) on variance**

- valid when $n$ iid data from dist whose domain no depend on $\theta$ and $\hat{\theta}$ unbiased.
- $V\left[\hat{\theta}\right] \geq \frac{1}{I_n(\theta)} = \frac{1}{n \cdot I(\theta)}$
- where the Fischer information $I(\theta) = -E\left[\frac{\delta^2}{\delta\theta^2} \log f_X(x \mid \theta)\right]$ (p_X for discrete)

**Central Limit Theorem**

- what if non-normal distributions but want infer about pop mean?
  - point estimates unaffected: don't need distr
  - CIs and HTs *are*, need sampl distrs.
- $\overline{X} \xrightarrow{d} Y \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- ex: $n = 100$ iid data from unknown dist with $\mu = 20, \sigma^2 = 4$. Find $P\left(19.8 \leq \overline{X} \leq 20.2\right)$
- ex: How many iid data do we need to draw from a dist with $\mu = 10, \sigma^2 = 2$ for $P\left(\overline{X} < 10.2\right) > 0.9$?

### **Confidence Intervals for norm dist parameters**

- review:
  - two sided CI: a random interval that fulfills $P\left(\hat{\theta}_L \leq \theta \leq \hat{\theta}_H\right) = 1 - \alpha$
  - coverage: $100(1-\alpha)\%$ of eval'd intervals overlap $\theta$
  - compute by: find statistic $Y$ with a $y_{\text{obs}}$. solve $F_Y(y_{\text{obs}} \mid \theta) = q$ for parameter ($\theta$), finding $q$ from the CI table

- now we must know how to uniroot it out

- ex: We draw $n = 100$ iid data from unknown distribution with mean $\mu$. have $\overline{x}_{\text{obs}} = 10, S^2 = 9$. find 95% two side CI for $\mu$.

  trick! invoke CLT and just treat $S$ as $\sigma$. proceed as expected with uniroot for compute step.

- ex: $n = 8$ iid data from normal with mean $\mu$, variance $\sigma^2$. have $s^2_{\text{obs}} = 6$. What is 90% upper bound on $\sigma^2$?

### **Hypothesis Tests again as well**

- review:
  - preconceived notion about dist param $\theta$ (e.g. normal mean). state that null hypothesis.
  - select a statistic $Y$ that informs $\theta$, write down sampl dist given null, and see if $y_{\text{obs}}$ is consistent with null sampl distr.

- if $y_{\text{obs}}$ falls in a rejection region, decide to reject null (otherwise, fail to reject)
  - $P(\text{reject null} \mid \text{null true}) = \alpha \longleftarrow$ user-set Type I error
  - $P(\text{fail to reject null} \mid \theta \text{ arb}) = \beta \longleftarrow$ Type II error (function of $\alpha, \theta$)
  - $P(\text{reject null} \mid \theta \text{ arb}) = 1 - \beta = \text{power}$

- Kolmorgorov-Smirnov (KS) test
  - $H_0$ : observed data are from some continuous distribution
  - or $H_0$ : two datasets are from same continuous distribution

Some details on the KS test are to be found in the book, but the gist is the following:

① The test statistic is $D_n = \sup \left| F_{X_n}(x) - F_X(x) \right|$

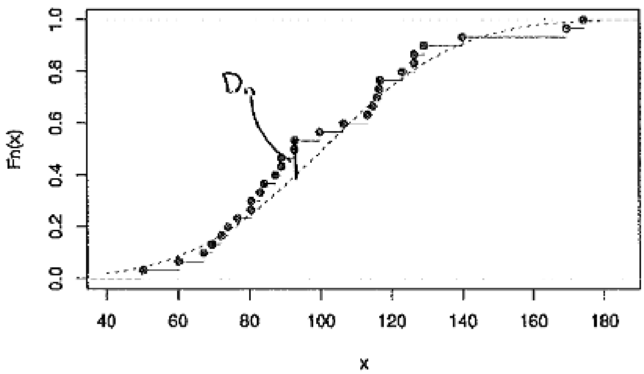
largest

empirical cdf of observed data (dots in plot)

cdf of specified distribution (curve in plot)

= "largest difference between what we observe and what we expect under the null"

② Under the null, $\sqrt{n}\, D_n$ is sampled from a Kolmogorov distribution




- Shapiro-Wilk test
  - stronger (more powerful) test
  - **only** for whether data are normally distributed
  - limited to $n \leq 5000$ data

**lecture on p-value, power, the normal mean**

- setting: let's say we want a hypothesis test about $\mu$ after get $n$ iid data (we assume/know normally distr)

- null/alt hypotheses:

  $H_0 : \mu = \mu_0$

  $H_A : \mu < \mu_0$ (lower-tail)

  $\mu \neq \mu_0$ (two-tail)

  $\mu > \mu_0$ (upper-tail)

- most common test statistic
  - $Y = \overline{X}$
  - where $E[Y] = \mu$, increases with $\mu$

- what is sampl dist for that statistic?
  - $\sigma^2$ known: $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ or $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
  - $\sigma^2$ unknown: $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$

- do what with these dists?
  - solve for rr boundaries: $F_{\overline{X}}(\overline{x}_{\text{RR}} \mid \mu_0) = q$
  - we fix $\mu_0$ so we solve via `qnorm()`, not `uniroot()`.
  - use table!

- do we need $\overline{x}_{\text{RR}}, t_{\text{RR}}$ to decide reject/not?
  - no! the $p$-value exists!
- $p$-value: prob that we observe $y_{\text{obs}}$ or "more extreme" statistic value, if $H_0$ is correct.
  - if $E[Y]$ incr with $\theta$, then ...
    - lower: $p = P(Y \leq y_{\text{obs}} \mid H_0)$
    - upper: $p = P(Y \geq y_{\text{obs}} \mid H_0)$
    - lower: $p = 2 \cdot \min[P(Y \leq y_{\text{obs}} \mid H_0), P(Y \geq y_{\text{obs}} \mid H_0)]$
    - can use cdf codes (e.g. `pnorm()`)
  - if $y_{\text{obs}} = y_{\text{RR}}$, then $p = \alpha$
  - if $H_0$ is correct, then $p \sim \text{Uniform}(0, 1)$, then $P(p \leq \alpha) = \int_0^\alpha dp = \alpha$ (think abt it! :3)
  - $p \neq$ prob that null correct! this makes no sense!
  - select $\alpha$ before $p$. don't $p$ hack.
- ex: $n = 9$ iid data from normal with mean $\mu$ and variance $\sigma^2 = 16$, $\overline{x}_{\text{obs}} = 11, \alpha = 0.05$. Find $p$ if $H_0 : \mu = \mu_0 = 10$ vs $H_A : \mu > \mu_0$
- test power: prob that we reject null given any (arb) value for $\theta$
  - $\text{power}(\theta) = P(\text{reject null} \mid \theta)$
  - implies $\text{power}(\theta = \theta_0) = \alpha$
  - $y_{\text{RR}}$ NOT needed to compute $p$ but IS needed to compute power.

**lecture on normal population variance**