EE325 Module 6

(Unofficial)

These slides are screenshots of the lectures

Viewers use it at your own risk

Module 6: Statistics
Parameter Estimation and Hypothesis Testing

D Manjunath

Networking Lab
Department of Electrical Engineering
Indian Institute of Technology

October 22, 2020

- Reiterating: Slides will be necessarily terse and you have to read from the textbook to get a full grasp of the material
- You are expected to take notes during the lecture.

Statistics

- Objective: Study a sample and systematically generalise to the *population* from whence the sample was obtained.
- Studying the whole population is impractical; only a part can be examined. This part is called the *sample*.
- Statistics is about making inferences from the sample to the population.

Statistics

- There are some numerical facts about the population that investigators want to know, e.g., average family income in a demographic.
- These numerical facts are called *parameters*.
- These cannot be determined exactly but can be *estimated* from *samples*.
- Major Issue: Acccuracy of the estimate.
  - How close is the estimate going to be to the true value?
  - How confident are we about this closeness? i.e., what is the probability that the estimate is 'wrong.'
  - Paramenters are estimated by statistics, or numbers which can be computed from the sample.
  - If $x_1, \ldots, x_n$ is the set of n measurements from n samples, some function $f(x_1, \ldots, x_n)$ is a statistic.
  - To summarise: Statistics are what the investigator knows from the samples. Parameters are what they want to know.

TRP Ratings

TRP में फर्जीवाड़ा: रिपब्लिक विवाद के बाद ब्रॉडकास्ट काउंसिल BARC ने न्यूज चैनलों की वीकली TRP लिस्ट पर अस्थायी रोक लगाई

मुंबई 6 दिन पहले

CBI Moves To Investigate Fake Ratings Case As Complaint Is Filed In UP

Republic TV has been demanding a CBI probe, accusing the Mumbai police of going after ' ' after their recent clash over the Sushant Singh Rajput case.

Reported by Arvind Gunasekar, Sreenivasan Jain. Edited by Deepshikha Ghosh | Updated: October 21, 2020 7:07 am IST

TRP scam case: BARC says disappointed with Republic TV for 'disclosing, misrepresenting' confidential communication

The statement comes after the Republic Media Network earlier in the day claimed that BARC, in an email said it had found no complaint or evidence of malpractice against Republic TV, Republic Bharat or any of its affiliates

The Manipulation of Television Rating Points: How TRPs work, the scam

Mumbai police are looking at alleged manipulation of Television Rating Points. How do channels score TRPs? What are the ways in which manipulation is possible, and how often has it been alleged in the past?

TRP ratings: background

- TV adspend is estimated at 27,000 crore rupees in 2020.
- Advertisers want to place their advertisements where they are most effective, i.e., choose slots where their target audience is present in required numbers and with required mindset to be 'programmed'.
- There are about 197 million TV homes in India, out of 298 million. About 55% of these in rural areas; a choice of about 850 channels.
- *Audience factor* is the number of people watching per TV. This is determined by the time of day and day of week and is obtained by different measures.
- *Reach* and *Target* audience—*Gross Rating Point* and *Target Rating Point* are used by advertisers to buy ad time.

TRP ratings: some methodology

- TAM is conducted in 45,000 homes that have cable TV service by placing "people meters" in these homes.
- The households are selected from class I towns (towns with population more than 1,00,000) around the country and are supposed to be representative of the TV viewing population.
- Some obvious problems:
  - The rural population is not covered at all! Rural viewing profile is very different from urban.
  - DTH is not covered.
- Data is collected overnight and the rating published on Friday.
- TRP rating determines the price of advertising in the slot.

TRP ratings: organisations involved

- In India, the popularity of different programs are measured with a number called the TRP (Televising Rating Point) rating. It is indicative of TV viewership of a program.
- "People Meters" are placed in a *sample* of the homes to determine what the households are watching.
- People meters essentially determines the channel to which the TV set is tuned. This information alongwith the times is recorded and stored suitably for later analysis.
- In India, TRP measurement is conducted by several organisations—Joint Industry Board Television Audience Measurement Research (JIB-TAM), ORG-MARG's Indian Television Audience Measurement (INTAM), (these two have now merged), Audience Measurement and Analytics (aMap), Doordarshan Audience Ratings team (DART);
- Broadcast Audience Research Council (BARC) is a cross industry body that oversees the TV audience measurement. Telecom Regulatory Authority of India (TRAI) regulates the process and provides guidelines for operation.

TRP ratings: issues

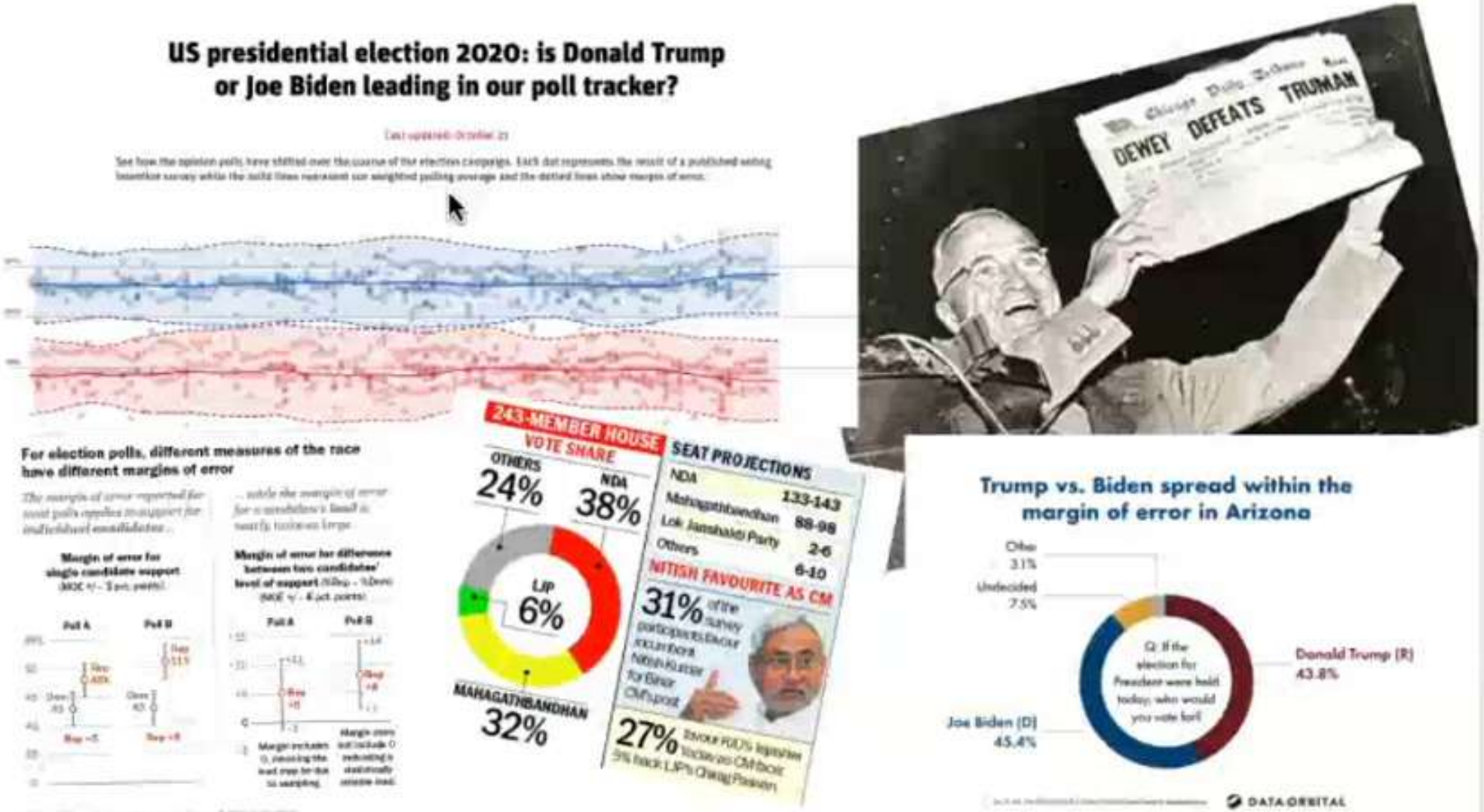
- An important question for advertiser: Do the same people watch the ad at different times?
- There is under reporting of cable homes because the service provider has to pay the broadcast company in proportion to the subscriber base.
- Channel availability is highly non uniform.
- Secrecy: The broadcasters and the cable TV suppliers should not know the families that are part of the survey.
- Effect of chasing TRP rating: Even the definition of news changes—entertainment and crime is more than 30% of the news!

TRP ratings: some methodology

- TAM is conducted in 45,000 homes that have cable TV service by placing "people meters" in these homes.
- The households are selected from class I towns (towns with population more than 1,00,000) around the country and are supposed to be representative of the TV viewing population.
- Some obvious problems:
  - The rural population is not covered at all! Rural viewing profile is very different from urban.
  - DTH is not covered.
- Data is collected overnight and the rating published on Friday.
- TRP rating determines the price of advertising in the slot.

TRP ratings: some more economics

- It costs more than twenty five crore rupees per year to obtain and analyse the data annually.
- Media planning and buying agencies use these ratings to place advertising time on TV shows.
- The buyers of the analysis may pay up to ten crore rupees per month.


US presidential election 2020: is Donald Trump or Joe Biden leading in our poll tracker?
See how the opinion polls have shifted over the course of the election campaign. Each dot represents the result of a published voting intention survey while the solid lines represent our weighted polling average and the dotted lines show margin of error.
For election polls, different measures of the race have different margins of error
The margin of error reported for most polls applies to support for individual candidates ...
... while the margin of error for a candidate's lead is nearly twice as large
Margin of error for single candidate support
Margin of error for difference between two candidates' level of support
Poll A
Poll B
PEW RESEARCH CENTER
243-MEMBER HOUSE
VOTE SHARE
OTHERS 24%
NDA 38%
LJP 6%
MAHAGATHBANDHAN 32%
SEAT PROJECTIONS
NDA 133-143
Mahagathbandhan 88-98
Lok Janshakti Party 2-6
Others 6-10
NITISH FAVOURITE AS CM
31% of the survey participants favour incumbent Nitish Kumar for Bihar CM's post
27% favour RJD's tejashwi Yadav as CM
5% back LJP's Chirag Paswan
DEWEY DEFEATS TRUMAN
Trump vs. Biden spread within the margin of error in Arizona
Other 3.1%
Undecided 7.5%
Q. If the election for President were held today, who would you vote for?
Donald Trump (R) 43.8%
Joe Biden (D) 45.4%
DATA ORBITAL

On Sampling

- Estimating the parameters from the sample is justified if the sample represents the population.
- This is impossible to check based on the measurements on the samples; to see whether sample looks like the population, we need to know the parameters which is what is being determined—a vicious circle.
- Thus the sampling methodology becomes important.
- The best methods involve the planned introduction of chance.
- How good can it get: In 2000, Gallup polled just 2014 people to determine that G W Bush will get 49% votes. He actually got 50.6%!
- And how bad: In the x1948 US Presidential election (Harry Truman vs Thomas Dewey) three separate polls said Dewey would get 50, 50 and 53 % of the vote and Truman would get 45, 44 and 38% of the vote. Actual result: 50% for Truman and 45% for Dewey!

Esimation

- Debugging a program: A program contains an unknown number of bugs, say N.
- Deliberately introduce n (known) bugs and see how many of these are caught, say n_1. Estimate N.
- Equivalent to calculating the volume of an arbitrary object N-dimensions with a membership oracle and a randomised algorithm.
- Other examples from before: Fish population in the Lake. n fish in the lake are marked. n are caught after some time and the number of marked fish is counted, say n_1. The experiment (catching the fish) is repeated K times with sample values $n_1, \ldots, n_K$. Estimate the number of fish in the Lake.

Estimators: Point and Interval Estimate

- Let $x = \{x_1, x_2, \ldots, x_n\}$ be sample measurements and the θ be the parameter of interest.
- x_i are realisations of random variables X_i.
- Any function of the sampeles x is called a *statistic*.
- $\hat{\theta} = g(x_1, x_2, \ldots, x_n)$ is a *point estimate* of θ.
- $\left(\mathsf{E}\left(\hat{\theta}\right) - \theta\right)$ is called the bias of the estimate.
- $\mathsf{E}\left(\left(\hat{\theta} - \theta\right)^2\right)$ is called the mean squared error of the estimate.
- An alterntaive to a point estimate is the *interval estimate* described using $\theta_l, \theta_u, \gamma$

$$\mathsf{Prob}(\theta_1 \leq \theta \leq \theta_u) \geq \gamma$$

- γ is called the *confidence level* and (θ_l, θ_u) is the the *interval estimate*.

Recap: Sums of Independent Random Variables

- Consider n the sum of n independent and identically distributed (iid) random variables—X_i, $i = 1, \ldots n$.
- Let $\mathsf{E}(\mathsf{X}_1) = \mu$ and $\mathsf{VAR}(\mathsf{X}_1) = \sigma^2$.
- Let the sample sum be $\mathsf{S}_n := \sum_{i=1}^{n} \mathsf{X}_i$.
- We know that $\mathsf{E}(\mathsf{S}_n) = n\mu$ and $\mathsf{VAR}(\mathsf{S}_n) = n\sigma^2$.
- From the Law of Large Numbers, $\frac{\mathsf{S}_n}{n} \longrightarrow \mu$,
  - Intuitively, if we take a large number of samples, the sample mean will 'approach the true mean.
- From Central Limit Theorem, with Z being the unit Normal,

$$\mathsf{Prob}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} < y\right) \rightarrow \mathsf{Prob}(\mathsf{Z} < y)$$

  - As n becomes large, the distribution of the approaches that of the Gaussian distribution with same mean and same variance as the sum.
  - Thus for large n, the Gaussian is a good approximation for the distribution of the sum of n iid random variables.

The Unit Normal

- Z is the zero-mean Gaussian random variable with unit variance.
- If X is a Gaussian random variable with mean μ and standard deviation σ, then $\frac{X-\mu}{\sigma}$ is unit normal.
- The above operation essentially *shifts and scales* the original Gaussian, it normalises it.
- Some 'unit normal' probabilities
  - $\text{Prob}(|Z| < 1) \approx 0.68$
  - $\text{Prob}(|Z| < 2) \approx 0.95$
  - $\text{Prob}(|Z| < 3) \approx 0.997$
  - $\text{Prob}(|Z| > 1) \approx 1 - 0.68 = 0.32$

The Unit Normal

- Prob$(0 \leq Z < 1) \approx$
- Prob$(0 \leq Z > 1) \approx$
- Prob$(0 \leq Z > 2) \approx$
- Prob$(Z > z_\alpha) = 1 - \alpha$

Percentile	90	95	97.5	99	99.5	99.9	99.95
α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
z_α	1.28	1.645	1.96	2.33	2.58	3.08	3.27

Sample Mean

- Assume a population with mean μ and variance σ^2.
  - μ is unknown and is to be estimated, σ is known.
  - Neither is known and are to be estimated.
- 'Obvious method': Obtain n independent samples whose values will be $X_1, \ldots X_n$ and calculate the sample mean as an estimator for the mean.

$$\text{Sample Mean} := \bar{X}_n := \frac{\sum_{i=1}^{n} X_i}{n}$$

- Clearly, $\bar{X}_n$ is a random variable. Characterising this random variable characterises the quality of the estimate.
- The expectation and variance of the sample mean are

$$E(\bar{X}_n) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sum_{i=1}^{n} E(X_i)}{n} = \mu$$

$$VAR(\bar{X}) = VAR\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sum_{i=1}^{n} VAR(X_i)}{n} = \frac{\sigma^2}{n}$$

- What is the distribution $\bar{X}_n$?
- How about 'an approximate distribution' of $\bar{X}_n$?

Sample Variance: Known mean

- Consider the sum of the square of the deviations from the mean, i.e.,

$$
\begin{aligned}
V &= \sum_{i=1}^{n} (X_i - \mu)^2 \\
\sum_{i=1}^{n} (X_i - \mu)^2 &= \sum_{i=1}^{n} X_i^2 - n\mu^2 \\
E(V) &= E\left(\sum_{i=1}^{n} (X_i - \mu)^2 \right) \\
&= \sum_{i=1}^{n} E\left((X_i - \mu)^2 \right) \\
&= n\sigma^2
\end{aligned}
$$

- Thus if μ is known then V/n is an unbiased estimator of $\mathsf{VAR}(X_1)$.

Sample Variance: Unknown mean

- Consider the sum of the square of the deviations from the mean, i.e.,

$$
\begin{aligned}
V &= \sum_{i=1}^{n} (X_i - \bar{X})^2 \\
\sum_{i=1}^{n} (X_i - \bar{X})^2 &= \sum_{i=1}^{n} X_i^2 - n(\bar{X})^2 \\
E(V) &= E\left(\sum_{i=1}^{n} X_i^2 - n(\bar{X})^2 \right) \\
&= \sum_{i=1}^{n} E(X_i^2) - nE((\bar{X})^2)\, E((\bar{X})^2) \\
&= nVAR(X_1) + n(E(X_1))^2 - nVAR(\bar{X}) - n(E(\bar{X})) \\
&= n\sigma^2 + n\mu^2 - n\left(\frac{\sigma^2}{n}\right) - n\mu^2 = (n-1)\sigma^2.
\end{aligned}
$$

- Thus, when μ is not known, $\frac{V}{n-1}$ is the unbiased estimate of $VAR(X_1)$.

Sampling from a Normal Population

- If X_i are i.i.d. Gaussian with mean μ and standard deviation σ then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

  is standard normal.

- Define

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

- $\bar{X}$ and S^2 are independent random variables.

Known Variance

- Since $\bar{X}_n$ is Gaussian, we can say

$$\text{Prob}\left(\mu - z_{1-\delta/2}\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + z_{1-\delta/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \frac{\delta}{2} - \frac{\delta}{2}$$

$$\text{Prob}\left(\bar{X}_n - z_{1-\delta/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\delta/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \delta = \gamma$$

- If the population is not from a normal distribution, invoke CLT to use the preceding as an approximation.
- Alternatively, use Chebyshev inequality to

$$\text{Prob}\left(\bar{X}_n - \frac{\sigma}{\sqrt{n\delta}} \leq \mu \leq \bar{X}_n + \frac{\sigma}{\sqrt{n\delta}}\right) \geq 1 - \delta = \gamma.$$

Distributions of $\bar{X}$ and S^2

- Can show that $\bar{X}$ and S^2 are independent with $\bar{X}$ being normal and $(n-1)S^2/\sigma^2$ being χ^2_{n-1}.
- This means $\sqrt{n}\frac{\bar{X}-\mu}{S}$ is Student t distribution, t_{n-1}.
  - $Z_1^2 + Z_2^2 + \cdots Z_n^2$ gives Chi-Square Random Variable with n degrees of freedom, denoted by χ^2_n.
  - $\frac{Z}{\sqrt{Y_n}}$ where Y is χ^2_n is Student t_n random variable.
- This helps in obtaining the confidence interval when the variance is unknown.

$$\text{Prob}\left(-t_{\alpha/2,n-1} < \sqrt{n}\frac{\bar{X}-\mu}{S} < t_{\alpha/2,n-1}\right) = 1-\alpha$$

- To obtain confidence interval for vairance note that $(n-1)S^2/\sigma^2$ is χ^2_{n-1}. Thus

$$\text{Prob}\left(\chi^2_{1-\alpha/2,n-1} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right) = 1-\alpha$$

Unknown Variance: Distributions of $\bar{X}$ and S^2

- X_i, $i = 1, \ldots, n$ are iid samples of Gaussian of mean μ and variance σ^2.
- Recall sample mean, $\bar{X}$, and sample variance, S^2, are

$$\begin{aligned} \bar{X} &= \frac{\sum_{i=1}^{n} X_i}{n} \\ S^2 &= \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} \end{aligned}$$

- Sum independent Gaussian random variables is also a Gaussian.
- Thus the sample mean $\bar{X}$, is Gaussian with mean μ and variance σ^2/n.
- Now consider

$$\begin{aligned} \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 + \frac{1}{\sigma^2} n(\bar{X} - \mu)^2 \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 + \left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \right)^2 \end{aligned}$$

Chi-Square and t Distributions

- If $Z, Z_1, \ldots Z_n$ are independent unit normal random variables then
  - $Z_1^2 + Z_2^2 + \cdots + Z_n^2$ is a χ^2 with n-degrees of freedom. It has mean 0 and variance $2n$.
  - The quantity

$$\frac{Z}{\sqrt{(Z_1^2 + Z_2^2 + \cdots + Z_n^2)/n}}$$

    has t-distribution. This has mean 0 and variance $n/(n-2)$.
- LHS is χ_n^2. Second term of RHS is χ_1^2. Hence reasonable to guess that first term of RHS is χ_{n-1}^2.

Likelihood Function

- $x_1, \ldots x_K$ is the set of measurements.
- Assume that the parameter set is θ.
- Question: What is the *likelihood* of observing of $x_1, \ldots x_K$ if the parameter value was θ?
  - Assume the population is Bernoulli with parameter p.
  - Assume the population is Poisson with parameter λ.
  - Assume the population is uniform in the interval $(0, a)$.
  - Assume the population is Gaussian with unit variance and mean μ.
  - Assume the population is exponential with parameter μ.

Maximum Likelihood Estimate

- Find the θ that maximises the likelihood of the observations.

$$\hat{\theta} = \arg\max_{\theta} L(x_1, \ldots, x_K|\theta)$$

- Can also maximise the Log Likelihood function.
- Likelihood and log likelihood functions and MLEs
- Bernouilli

$$\begin{aligned} L(\underline{x}; p) &= \prod_{i=1}^{K} p^{x_i}(1-p)^{1-x_i} = p^n(1-p)^{K-n} \\ \hat{p} &= \frac{\sum_{i=1}^{K} x_i}{K} \end{aligned}$$

- Poisson

$$\begin{aligned} L(\underline{x}; \lambda) &= \prod_{i=1}^{K} e^{-\lambda}\frac{\lambda x_i}{x_i!} = e^{-K\lambda}\frac{\lambda^{\sum_{i=1}^{K} x_i}}{\prod_{i=1}^{K} x_i} \\ \log L(\underline{x}; \lambda) &= -K\lambda + \sum_{i=1}^{K} x_i \log\lambda - \log(\prod_{i=1}^{K} x_i) \\ \hat{\lambda} &= \frac{\sum_{i=1}^{K} x_i}{K} \end{aligned}$$

Maximum Likelihood Estimate

- Gaussian with known σ and unknown μ.

$$
\begin{aligned}
L(\underline{x};\lambda) &= \prod_{i=1}^{K} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(\sqrt{2\pi}\sigma)^K} e^{-\sum_{i=1}^{K}\frac{(x_i-\mu)^2}{2\sigma^2}} \\
\log L(\underline{x};\lambda) &= -K\log(\sqrt{2\pi}\sigma) - \sum_{i=1}^{K}\frac{(x_i-\mu)^2}{2\sigma^2} \\
\hat{\mu} &= \frac{\sum_{i=1}^{K} x_i}{K}
\end{aligned}
$$

- MLE for Uniform in $(0, a)$

$$
\begin{aligned}
L(\underline{x};\lambda) &= \frac{1}{a^K} \\
\hat{a} &= \max\{x_1, x_2, \ldots, x_K\}
\end{aligned}
$$

- Expectation of Estimate is the **bias** and variance is the mean squared error.

Minimum Mean Square Error Estimate

- $\hat{\theta} = \arg\min_{\theta} \left(\mathsf{E}\left(\left(\theta - \hat{\theta} \right)^2 \right) \right).$

Hypothesis Testing

- Not interested in explicitly estimating the parameters but in verifying a statistical hypotheses about them.
  - value of single parameter: (1) $\mu = 1$, (2) $\mu > 1$
  - values of a set of parameters: (1) $\mu = 10$, $\sigma > 1$
  - distribution of the population: uniform, normal, exponential, etc.
- It is a hypothesis because it is not known if it is true.
- Test method is as follows: Design a suitable random experiment and accept the hypothesis if the data from the experiment is *consistent* with the hypothesis.
- The notion of *consistent* needs to be made more precise.
- Accepting the hypothesis does not mean "claiming that the hypothesis is true;
- Rather, we are saying *experimental data appears to be consistent with the hypothesis*. Once again notion of 'appears to be' will be made more precise.

Hypothesis Testing

- **Null Hypothesis,** denoted by H_0 is to state that the initial claim or prior belief is true.
  - If there is the claim of a better drug, H_0 will be to state that this is not better than the current drug. (cures a higher fraction, or has lesser side effects)
  - If a new vendor to procure components is being tried, the new vendor is not any better than the current. (rejection ratio of the components is not lower).
  - If a new curriculum is being tested, then ....
- There is insufficient evidence in the data to change our beliefs or the 'status quo.'
- The word null means "of no value, effect or consequence," which suggests that H_0 should be identified with the hypothesis of no change, no effect or no improvement.
- H_0 is tested against an alternate hypothesis H_a or H_1.

Test Procedures

- Conduct the random experiment and obtain the samples.
- Define the **Test Statistic,** a function of the sample data on which the decision to accept or reject H_0 is based.
- Define a **Rejection Region,** the set of all test statistic values for which H_0 will be rejected.
- Immediate question: How to obtain the rejection region? We need to see what are the possible errors from our experiment and conclusion.
- **Type I error:** Rejecting H_0 when H_0 is indeed true. The probability of Type I error is usually denoted by α.
- **Type II error:** Not rejecting H_0 when H_0 is false. The probability of Type II error is usually denoted by β.

Example I

- A new curriculum will make our students happier. Currently, 25% of our students are satisfied with the curriculum.
- You have resources for 20 samples. Let p be the actual fraction of students that are satisfied with the new curriculum.
- $H_0 : p = 0.25$
  $H_1 : p > 0.25$
- Let N be the number of students satisfied with the new curriculum.
- Consider the rejection region $N \geq 8$.
- If H_0 is true then N is a binomial random variable with parameters 20 and 0.25.
- Thus α is the probability of a binomial random variable with parameters $(20, 0.25)$ being greater than 8—$\alpha = 0.102$.
- The value of β depends on the true value of p.
- With $p = 0.3$, β is the probability of $N \leq 7$ for a binomial random variable with parameters $(20, 0.3)$. $\beta = 0.772$.

Example

- The monthly expenditure on food of a household income is Gaussian with mean 7,500 and standard deviation 900. With large retail stores (supermarkets in malls), the mean expenditure will reduce but the standard deviation will remain the same. We will check the expenditure of 25 households.
- Let X be the expenses after the supermarkets enter the area.
- $H_0 : X$ is Gaussian with $\mu = 7500$ and $\sigma = 900$.
  $H_1 : X$ is Gaussian with $\mu < 7500$ and $\sigma = 900$
- Test statistic is the sample mean, $\bar{X}$, of the expenses. This has $\sigma = 900/5 = 180$.
- Consider the rejection region $\bar{X} < 7080$.
- $\alpha =$ is the probability of a normal random variable with parameters $(7500, 180)$ being lesser than 7080—$\alpha = 0.01$.
- The value of β depends on the true value of μ.
- With $\mu = 7200$, β is the probability of $\bar{X} \geq 7200$ for a normal variable with parameters $(7200, 900)$. $\beta = 0.7486$.
- With $\mu = 7000$, $\beta = 0.174$.

- In example 1, change rejection region to ≥ 9.
  $\alpha = 0.041$, $\beta(0.3) = 0.887$, and $\beta(0.5) = 0.252$.
- In example 2, change rejection region to ≤ 7200.
  $\alpha = 0.05$, $\beta(7200) = 0.5$, and $\beta(7000) = 0.1335$.
- Summary: Decreasing size of rejection region to obtain a smaller value of α results in a larger value of β for a particular set of the parameters.

Some more terminology

- Simple hypothesis: If true, H_0 completely specifies the population. Both the preceding examples are simple hypotheses.
- Composite hypothesis: Not a simple a hypothesis.
- *Rejection Region:* The set of outcomes for which H_0 is rejected.
- Probability of a Type I error, α, is also called the significance of the test.
- Probability of a Type II error depends on true value of parameter, i.e., β is a function of p and μ in the preceding examples.
- **Important: Objective is not to explicitly determine if H_0 is true. Rather it is to determine if its validity is consistent with experimental data.** x
- H_0 is rejected if experimental data are very unlikely if H_0 is true.
- $1 - \beta(p)$ (or $1 - \beta(\mu)$) is called the *power function* of the test.

Testing for Mean of a Normal Population with Known Variance

- $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.
- Test obtains $X_1, \ldots, X_n$ and computes the sample $\bar{X}$.
- A reasonable rejection region: $\{X_1, \ldots, X_n : |\bar{X} - \mu_0| \geq c\}$
- $\bar{X}$ is Gaussian with unknown mean μ and known standard deviation σ.
- If α is specified, choose c such that $\text{Prob}(|\bar{X} - \mu_0| > c \mid \mu = \mu_0) = \alpha$;
  Let $Z := \frac{\bar{X} - \mu_0}{(\sigma/\sqrt{n})}$

$$
\begin{aligned}
\text{Prob}(|\bar{X} - \mu_0| > c \mid \mu = \mu_0) &= \alpha; \\
\text{Prob}\left(|Z| > \frac{c\sqrt{n}}{\sigma}\right) &= \alpha \\
2\text{Prob}\left(Z > \frac{c\sqrt{n}}{\sigma}\right) &= \alpha \\
\text{Prob}(Z > z_{\alpha/2}) &= \alpha/2 \\
\frac{c\sqrt{n}}{\sigma} &= z_{\alpha/2} \\
c &= \frac{z_{\alpha/2}\sigma}{\sqrt{n}}
\end{aligned}
$$

Testing for Mean of a Normal Population with Known Variance

- Reject H_0 if $\frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| > z_{\alpha/2}$
  Accept H_0 if $\frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| \leq z_{\alpha/2}$
- A reasonable choice of α depends on the specifics of the situation, essentially the 'cost' of a Type I error. It will also depend on 'the conviction' about H_0. Strong beliefs would require strong data to 'overthrow' H_0.
- An alternative view:
  - Consider $v = \frac{\bar{X} - \mu_0}{(\sigma/\sqrt{n})}$
  - Let p be the probability that the unit normal exceeds v.
  - Now observe that p is *critical significance level* of the test, i.e, if $\alpha < p$, H_0 is accepted.
  - p is called the p-value of the test.
  - p is the probability, caclulated assuming that H_0 is true, of obtaining the value of the test statistic (e.g., sample mean) at least as contradictory to H_0 as the value calculated from the available sample.

Example

- H_0 : X is normal with mean $\mu = 8$ and $\sigma = 2$
  H_1 : X is normal with mean $\mu \neq 8$ and $\sigma = 2$
- From five samples, $\bar{X} = 9.5$.
- Test statistic:
$$\frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| = \frac{\sqrt{5}}{2} \times 1.5$$
- For 5% level of significance, i.e., $\alpha = 0.05$, $z_{0.025} = 1.96$.
- Since test statistic is less than $z_{\alpha/2}$ accept H_0 at 5% significance.
- If 'loosen' the criterion, i.e. choose $\alpha = 0.1$, then $z_{0.05} = 1.645$ and H_0 is rejected.
- If $\bar{X} = 8.5$, then $v = 0.559$.
  $\mathsf{Prob}(|Z| > 0.559) = 2\mathsf{Prob}(Z > 0.559) = 2 \times 0.288 = 0.576$.
- For $\bar{X} = 11.5$, $p = 0.00005$.

Remarks

- For tests involving mean of a normal population with known variance, we use the sample mean from n iid samples.
- The test statistic $Z = \frac{\bar{X}-\mu_0}{(\sigma/\sqrt{n})}$ measures the deviation of the sample mean from μ_0 in "units of the standard deviation."
- H_0 is rejected if Z is sufficiently large in the direction pointed to by H_1.
- Connections can be drawn with "Interval Estimate" of the mean of a normal population.

$$\text{Prob}\left(\mu \in \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\,,\, \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\right) \;=\; 1-\alpha$$

- Hence if $\mu = \mu_0$, then with probability $(1-\alpha)$ it will be in the interval $\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\,,\, \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$.
- This means that H_0 is rejected with significance level α if

$$\mu_0 \;\notin\; \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\,,\, \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

Type II Error

- We now obtain $\beta(\mu)$.
- Define $Z = \frac{\bar{X}-\mu}{(\sigma/\sqrt{n})}$. Of course Z is the unit normal if the $\bar{X}$ were Gaussian with mean μ and variance σ^2/m.

$$
\begin{aligned}
\beta(\mu) &= \text{Prob}\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) \\
&= \text{Prob}\left(-z_{\alpha/2} - \frac{\mu}{\sigma/\sqrt{n}} \leq \frac{\bar{X}-\mu_0-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} - \frac{\mu}{\sigma/\sqrt{n}}\right) \\
&= \text{Prob}\left(-z_{\alpha/2} - \frac{\mu}{\sigma/\sqrt{n}} \leq Z - \frac{\mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2} - \frac{\mu}{\sigma/\sqrt{n}}\right) \\
&= \text{Prob}\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}} - z_{\alpha/2} \leq Z \leq \frac{\mu_0-\mu}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) \\
&= \Phi\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) - \Phi\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}} - z_{\alpha/2}\right)
\end{aligned}
$$

Type II Error and Number of Samples

- Rejection region is determined by α.
- $\beta(\mu)$ can be used to determine the number of samples that may be required.
- Assume a requirement that $\beta(\mu_1) = \beta$, i.e.,

$$\Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{\alpha/2}\right) = \beta.$$

- Let $\mu_1 > \mu_0$; then

$$\frac{\mu_0 - \mu_1}{(\sigma/\sqrt{n})} - z_{\alpha/2} \leq -z_{\alpha/2}$$

- Since Φ is an increasing function; hence

$$\Phi\left(\frac{\mu_0 - \mu_1}{(\sigma/\sqrt{n})} - z_{\alpha/2}\right) \leq \Phi(-z_{\alpha/2}) = \alpha/2 \approx 0$$

$$\beta \approx \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) = \Phi(z_{-\beta})$$

$$-z_\beta = \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{\alpha/2}$$

- Solve for n.

Bernoulli Populations

- H_0: $p \leq p_0$ versus H_1: $p > p_0$.
- Random binary valued samples X_i. Sample statistic $X = \sum_{i=1}^{n} X_i$.

$$\text{Prob}(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

- This is an increasing function in p. And if H_0 is true, then

$$\text{Prob}(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}$$

- Reasonable rejection region would be $X > K^*$ where

$$K^* = \min\{k : \sum_{i=k}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} < \alpha\}$$

- It is easier to determine the p-value first. Let $X = x$.

$$p - \text{value} = \sum_{i=x}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}$$

Accept for $\alpha < p$-value.

Example

- A manufacturer wants to give a six year warranty on a product. Expectation is that 90% will last longer than six years. "Accelerated life testing" is performed on 20 samples and results obtained. Let X be the number that survived.
- H_0: $p = 0.9$ versus H_1: $p_1 < 0.9$
- If the desired significance is $\alpha = 0.05$, then we want

$$K^* = \min\{k : \sum_{i=k}^{20} \binom{20}{i} 0.9^i (1 - 0.9)^{20-i} < \alpha\}$$

- Can be seen that $K^* = 15$ and $\alpha = 0.043$.
- HW: Evaluate β for $p = 0.8$.

Bernoulli Populations: Large Samples

- X/n is an unbiased estimator for p.
- And X is approximately normal with mean np and variance $np(1-p)$.
- Thus $\frac{X-np}{\sqrt{np(1-p)}}$ is approximately unit normal.
- And all the theory from the previous "part" applies!
- For example, reject H_0 if

$$\frac{X - np_0}{\sqrt{np_0(1-p_0)}} \geq z_\alpha$$

- What do we do when H_0 is simple. i.e.. H_0: $p = p_0$ versus H_1: $p \neq p_0$?
- Homework!

Testing Two Populations

- Motivation:
  - Two candidate treatments for a disease; interested in fraction cured.
  - Two processes to manufacture a product; interested in fraction of defects.
  - Or closer to you—two coaching classes; interested in fraction of successes.
- Probability of success is p_i for population i.
- With independent sampling all successes are independent with appropriate probability.
- Unequal sample sizes—n_i from population i; and X_i successes from n_i samples of population i.
- X_i is a Binomial random variable with parameters (n_i, p_i).
- H_0: $p_1 = p_2$ versus H_1: $p_1 \neq p_2$.

Testing Two Populations

- Also under H_0 the samples form a population of size $(n_1 + n_2)$. And the total number of successes is Binomial with parameters $(n_1 + n_2, p_1)$.
- Let $X_1 + X_2 = k$.
- Note that we do not know p_1 or p_2; we are just checking to see if they are equal.
- Thus, under H_0, we expect X_1/n_1 to be close to X_2/n_2.
- Obtain the conditional probability of X_1, given k

$$\text{Prob}(X_1 = i | X_1 + X_k) = \frac{\binom{n_1}{i}\binom{n_2}{k-i}}{\binom{n_1+n_2}{k}}$$

- This is hypergeometric with parameters (n_1, n_2, k).
- Reject H_0 if X_1 is too small or too large.
- Assume the test resulted in $X_1 = x_1$.
- Reject H_0 if $\text{Prob}(X \leq x_1) \leq \alpha/2$ or $\text{Prob}(X \geq x_1) \leq \alpha/2$ where X is hypergeometric with parameters (n_1, n_2, k).
- Homework: What is the p-value of the test.
- This is called the Fisher-Irwin test.