Week 13 notes

Deviation and Large Numbers

01 Theory - Sample mean

Sample mean and its variance

The **sample mean** of a set $X_1, X_2, \dots$ of IID random variables is an RV that averages the first n instances:

$$M_n(X) \quad = \quad \frac{1}{n}\Big(X_1 + \cdots + X_n\Big)$$

Statistics of the sample mean (for any i):

$$E[M_n(X)] \;=\; E[X_i] \qquad\qquad \mathrm{Var}[M_n(X)] \;=\; \frac{\mathrm{Var}[X_i]}{n}$$

The sample mean is typically applied to repeated trials of an experiment. The trials are independent, and the probability distribution of outcomes should be identical from trial to trial.

Notice that the variance of the sample mean limits to 0 as $n \to \infty$. As more trials are repeated, and the average of all results is taken, the fluctuations of this average will shrink toward zero.

As $n \to \infty$ the *distribution* of $M_n(X)$ will converge to a PMF with all the probability concentrated at $E[X_i]$ and none elsewhere.

02 Theory - Tail estimation

Every distribution must trail off to zero for large enough $|X|$. The regions where X trails off to zero (large magnitude of X) are informally called 'tails'.

Tail probabilities

A **tail probability** is a probability with one of these forms:

$$P\big[\,X \geq c\,\big] \qquad P\big[\,X \leq -c\,\big] \qquad P\big[\,|X - \mu_X| \geq c\,\big]$$

Markov's inequality

Assume that $X \geq 0$. Take any $c > 0$.

Then the **Markov's inequality** states:

$$P\big[\,X \geq c\,\big] \quad \leq \quad \frac{\mu_X}{c}$$

Chebyshev's inequality

Take any X, and $c > 0$.

Then the **Chebyshev's inequality** states:

$$P\big[\,|X-\mu_X| \geq c\,\big] \quad \leq \quad \frac{\sigma_X^2}{c^2}$$

Markov vs. Chebyshev

Chebyshev's inequality works for any X, and it usually gives a better estimate than Markov's inequality.

The main value of Markov's inequality is that it only requires knowledge of μ_X.

Think of Chebyshev's inequality as a tightening of Markov's inequality using the additional data of σ_X.

Derivation of Markov's inequality - Continuous RV

Under the hypothesis that $X \geq 0$ and $c > 0$, we have:

$$\mu_X = E[x] \;=\; \int_0^\infty x f_X(x)\,dx \quad = \quad \int_0^c x f_X(x)\,dx + \int_c^\infty x f_X(x)\,dx$$

On the range $c \leq x < \infty$ we may convert x to c, making the integrand bigger:

$$\int_c^\infty x f_X(x)\,dx \quad \geq \quad \int_c^\infty c f_X(x)\,dx$$

Simplify:

$$\int_c^\infty c f_X(x)\,dx \ggg \quad c\int_c^\infty f_X(x)\,dx \quad \ggg \quad cP\big[\,X \geq c\,\big]$$

Also:

$$\int_0^c x f_X(x)\,dx \quad \geq \quad 0$$

Therefore:

$$\int_0^\infty x f_X(x)\,dx \quad \geq \quad cP\big[\,X \geq c\,\big]$$

$$\ggg \quad E[x] \quad \geq \quad cP\big[\,X \geq c\,\big]$$

Extra - Derivation of Chebyshev's inequality

Notice that the variable $(X-\mu_X)^2$ is always positive. Chebyshev's inequality is a simple application of Markov's inequality to this variable.

Specifically, using c^2 as the Markov constant, Markov's inequality yields:

$$P\left[\,(X-\mu_X)^2 \geq c^2\,\right] \quad \leq \quad \frac{E\left[\,(X-\mu_X)^2\,\right]}{c^2}$$

Then, by monotonicity of square roots:

$$(X-\mu_X)^2 \geq c^2 \quad \Longleftrightarrow \quad |X-\mu_X| \geq c$$

And of course $E\left[\,(X-\mu_X)^2\,\right] = \sigma_X^2$. Chebyshev's inequality follows.

03 Illustration

Markov's inequality derivation - Discrete RV

Derive Markov's inequality for a discrete RV.

Example - Markov and Chebyshev

A tire shop has 500 customers per day on average.

(a) Estimate the odds that more than 700 customers arrive today.

(b) Assume the variance in daily customers is 10. Repeat (a) with this information.

Solution

Write X for the number of daily customers.

(a) Using Markov's inequality with $c = 700$, we have:

$$P\left[\,X \geq 700\,\right] \quad \leq \quad \frac{500}{700} \approx 0.71$$

(b) Using Chebyshev's inequality with $c = 200$, we have:

$$P\left[\,|X-500| \geq 200\,\right] \quad \leq \quad \frac{100}{200^2} \approx 0.0025$$

The Chebyshev estimate is much smaller!

04 Theory - Law of Large Numbers

Let $X_1, X_2, \ldots$ be a collection of IID random variables with $\mu = E[X_i]$ for any i, and $\sigma^2 = \mathrm{Var}[X_i]$ for any i.

Recall the sample mean:

$$M_n(X) \quad = \quad \frac{1}{n}\Big(X_1 + \cdots + X_n\Big)$$

Recall that $\mathrm{Var}[M_n(X)] = \frac{\sigma^2}{n}$.

Law of Large Numbers (weak form)

For any $\varepsilon > 0$, by Chebyshev's inequality we have:

$$P\big[\,|M_n(X)-\mu| \geq \varepsilon\,\big] \quad \leq \quad \frac{\sigma^2}{n\varepsilon^2} \qquad \text{(finite LLN)}$$

Therefore:

$$\lim_{n\to\infty} P\big[\,|M_n(X)-\mu| \geq \varepsilon\,\big] \quad = \quad 0$$

And the complement:

$$\lim_{n\to\infty} P\big[\,|M_n(X)-\mu| < \varepsilon\,\big] \quad = \quad 1 \qquad \text{(infinite LLN)}$$

05 Illustration

Example - LLN: Average winnings

A roulette player bets as follows: he wins \$100 with probability 0.48 and loses \$100 with probability 0.52. The expected winnings after a single round is therefore $\$100 \cdot 0.48 - \$100 \cdot 0.52$ which equals $-\$4$.

By the LLN, if the player plays repeatedly for a long time, he expects to lose \$4 per round on average.

The 'expects' in the last sentence means: the PMF of the cumulative average winnings approaches this PMF:

$$P_{M_n(X)}(k) = \begin{cases} 1 & k = \$4 \\ 0 & k \neq \$4 \end{cases}$$

This is by contrast to the 'expects' of expected value: the probability of achieving the expected value (or something near) may be low or zero! For example, a single round of this game.

Exercise - Enough samples

Suppose $X_1, X_2, \ldots$ are IID samples of $X \sim \text{Ber}(0.6)$.

(a) Compute $E[X]$ and $\text{Var}[X]$ and $\text{Var}\,[M_{100}(X)]$.

(b) Use the finite LLN to find α such that:

$$P\big[\,|M_{100}(X) - 0.6| \geq 0.05\,\big] \leq \alpha$$

(c) How many samples n are needed that to guarantee that:

$$P\big[\,|M_n(X) - 0.6| \geq 0.1\,\big] \leq 0.05$$

Statistical testing

06 Theory - Significance testing

Significance test

Ingredients of a significance test (unary hypothesis test):

- Null hypothesis event H_0
  - Identify a Claim
  - Then: H_0 is background assumption (supposing Claim isn't known)
  - Goal is to *invalidate H_0* in favor of Claim
- Rejection Region (decision rule): an event R
  - R is *unlikely* assuming H_0
  - Directionality: R is *more likely* if Claim
  - Write R in terms of **decision statistic** X and **significance level** α
- Ability to compute $P[R \mid H_0]$
  - Usually: inferred from $f_{X|H_0}$ or $P_{X|H_0}$
  - Adjust R to achieve $P[R \mid H_0] = \alpha$

Significance level

Suppose we are given a null hypothesis H_0 and a rejection region R.

The **significance level of R** is:

$$\begin{aligned} \alpha &= P\big[\text{reject } H_0 \mid H_0 \text{ is true}\big] \\ &= P[R \mid H_0] \end{aligned}$$

Sometimes the condition is dropped and we write $\alpha = P[R]$, e.g. when a background model without assuming H_0 is not known.

Null hypothesis implies a distribution

Frequently H_0 will *not* take the form of an event in a sample space, $H_0 \subset S$.

Usually S is unspecified, yet H_0 determines a known distribution.

At a minimum, the assumption of H_0 must determine numbers $P[R \mid H_0]$.

More generally, we do **not** need these details:

- Background sample space S
- Non-conditional distribution (full model): f_X or P_X
- Complement conditionals: $f_{X|H_0^c}$ or $P_{X|H_0^c}$

In basic statistical inference theory, there are two kinds of error.

- Type I error concludes with rejecting H_0 when H_0 is true.
- Type II error concludes with maintaining H_0 when H_0 is false.

Type I error is usually a bigger problem. We want to consider H_0 "innocent until proven guilty."

	H_0 **is true**	H_0 **is false**
Maintain null hypothesis	Made right call	Wrong acceptance Type II Error
Reject null hypothesis	Wrong rejection Type I Error	Made right call

---

To *design a significance test at α,* we must identify H_0, and specify R with the property that $P[R \mid H_0] = \alpha$.

When R is written using a variable X, we must choose between:

- One-tail rejection region: x with $R(x) \leq r$ or x with $R(x) \geq r$
- Two-tail rejection region: x with $|R(x) - \mu| \geq c$

07 Illustration

Example - One-tail test: Weighted die

Your friend gives you a single regular die, and say she is worried that it has been weighted to prefer the outcome of 2. She wants you to test it.

Design a significance test for the data of 20 rolls of the die to determine whether the die is weighted. Use significance level $\alpha = 0.05$.

Solution

Let X count the number of 2s that come up.

The Claim: "the die is weighted to prefer 2"
The null hypothesis H_0: "the die is normal"

Assuming H_0 is true, then $X \sim \text{Bin}(20, 1/6)$, and therefore:

$$P_{X|H_0}(k) \quad = \quad \binom{20}{k}(1/6)^k(5/6)^{20-k}$$

⚠️ Notice that "prefer 2" implies the claim is for *more* 2s than normal.

Therefore: Choose a one-tail rejection set.

Need r such that $P[X \geq r \mid H_0] = 0.05$

- Equivalently: $P[X < r \mid H_0] = 0.95$

Solve for r by computing conditional CDF values:

$k:$	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
$F_{X\mid H_0}(k):$	0.026	0.130	0.329	0.567	0.769	0.898	0.963	0.989

Therefore, choose $r = 6$. Then $P[X \geq r \mid H_0] < 0.04$ and no smaller (integer) r will produce significance below 0.05.

The final answer is:

$$\boxed{R \;=\; \{x \mid x \geq 6\}}$$

Two-tail test: Circuit voltage

A boosted AC circuit is supposed to maintain an average voltage of $130\,\mathrm{V}$ with a standard deviation of $2.1\,\mathrm{V}$. Nothing else is known about the voltage distribution.

Design a two-tail test incorporating the data of 40 independent measurements to determine if the expected value of the voltage is truly $130\,\mathrm{V}$. Use $\alpha = 0.02$.

Solution

Use $M_{40}(V)$ as the decision statistic, i.e. the sample mean of 40 measurements of V .

The Claim to test: μ is not 130
The null hypothesis H_0: $\mu = 130$

Rejection region:

$$|M_{40} - 130| \geq c$$

where c is chosen so that $P\big[\,|M_{40} - 130| \geq c\,\big] = 0.02$

Assuming H_0, we expect that:

$$E[M_{40}] = 130 \qquad \sigma^2 = \mathrm{Var}[M_{40}] = \frac{2.1^2}{40} \approx 0.110$$

Recall Chebyshev's inequality:

$$P\big[\,|M_{40} - 130| \geq c\,\big] \leq \frac{\sigma^2}{c^2} \approx \frac{0.110}{c^2}$$

Now solve:

$$\frac{0.110}{c^2} = 0.2 \quad \gg\gg \quad c \approx 0.74$$

Therefore the rejection region should be:

$$M_{40} < 129.26 \quad \cup \quad 130.74 < M_{40}$$

☰ One-tail test with a Gaussian: Weight loss drug

Assume that in the background population in a specific demographic, the distribution of a person's weight W satisfies $W \sim \mathcal{N}(190, 24)$. Suppose that a pharmaceutical company has developed a weight-loss drug and plans to test it on a group of 64 individuals.

Design a test at the $\alpha = 0.01$ significance level to determine whether the drug is effective.

Solution

Since the drug is tested on 64 individuals, we use the sample mean $M_{64}(W)$ as the decision statistic.

The Claim: "the drug is effective in reducing weight"
The null hypothesis H_0: "no effect: weights on the drug still follow $\mathcal{N}(190, 24)$"

Assuming H_0 is true, then $W \sim \mathcal{N}(190, 24)$.

⚠️ One-tail test because the drug is expected to *reduce* weight (unidirectional).

Rejection region:

$$M_{64}(W) \leq r$$

Compute $\frac{24}{\sqrt{64}} = 3$.

Since $W \sim \mathcal{N}(190, 24)$, we know that $M_{64}(W) \sim \mathcal{N}(190, 3^2)$.

Furthermore:

$$\frac{M_{64}(W) - 190}{3} \quad \sim \quad \mathcal{N}(0, 1)$$

Then:

$$\begin{aligned} P[M_{64}(W) < r] \quad &= \quad P\left[Z < \frac{r - 190}{3}\right] \\ &= \quad \Phi\left(\frac{r - 190}{3}\right) \end{aligned}$$

Solve:

$$P[M_{64}(W) < r] = 0.01$$

$$\gg\gg \quad \Phi\left(\frac{r-190}{3}\right) = 0.01$$

$$\gg\gg \quad \Phi\left(\frac{190-r}{3}\right) = 0.99$$

$$\gg\gg \quad \frac{190-r}{3} = 2.33$$

$$\gg\gg \quad r = 183.01$$

Therefore, the rejection region:

$$M_{64}(W) \leq 183.01$$