Lecture 9: Large Random Samples

- Linear Combinations
- Markov and Chebyshev Inequalities
- Law of Large Numbers
- Variability in Estimates
- Central Limit Theorem

Introduction

– We are now set to round up our discussions on probability. We will still talk about probability today but in the context of statistics and how it helps with inference.

– From the next lecture, we will move to statistical inference.

– For today, we will learn about properties of linear combinations and the behavior of large random samples.

Linear Combinations

A **linear combination** of random variables $X_1, \ldots, X_n$ is a new random variable Y such that

$$Y = a_1 X_1 + \cdots a_n X_n = \sum_{i=1}^{n} a_i X_i$$

where the a_i's are known constants.

Some important linear combinations include:

– The sample mean, $\overline{X}$, in which each a_i equals $1/n$.

– A difference, $X_1 - X_2$, in which $a_1 = 1$ and $a_2 = -1$. This is helpful when deciding whether, say, one brand of lightbulb outlasts another brand, or whether one company outperforms another.

Expectation and Variance of Linear Combinations

Let X_i have mean μ_i and variance σ_i^2. Then

$$\mathbb{E}[Y] = \mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i] = \sum_{i=1}^{n} a_i \mu_i.$$

This holds even when the X_i's are dependent. It follows because integration (or sum) is a linear operator: $\int a_i x g_i(x)\, dx = a_i \int x g_i(x)\, dx = a_i \mu_i$. Also,

$$\mathbb{V}[Y] = \mathbb{V}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \text{Cov}[X_i, X_j]$$

Why? $\mathbb{V}[Y] = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2$ and $Y^2 = (a_1 X_1 + \ldots + a_n X_n)^2$ which generates the cross-product terms that define $\text{Cov}[X_i, X_j]$. It takes some algebra but you should work it out to convince yourself.

Expectation and Variance of Linear Combinations (Cont'd)

– Looking at the definitions of variance and covariance, we can see that $\text{Cov}[X_i, X_i]$ is just the variance σ_i^2. So we can write

$$\mathbb{V}[Y] = \sum_{i=1}^{n} a_i^2 \sigma_i^2 + 2 \sum_{i<j} a_i a_j \text{Cov}[X_i, X_j].$$

– In the special case when the random variables are independent, then the covariances are all zero and this simplifies to

$$\mathbb{V}[Y] = \sum_{i=1}^{n} a_i^2 \sigma_i^2.$$

Markov and Chebyshev Inequalities

Markov Inequality (D.S. Theorem 6.2.1): Let X be a random variable such that $\mathbb{P}(X \geq 0) = 1$. Then for every real number $t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}$$

Chebyshev Inequality (D.S. Theorem 6.2.2): Let X be a random variable for which Var(X) exists. Then for every number $t > 0$,

$$\mathbb{P}(|X - E(X)| \geq t) \leq \frac{\mathbb{V}(X)}{t^2}$$

These two inequalities are extremely useful in practice, since they **do not require knowing the exact distribution** of the random variable.

Obviously, they only give us upper bounds (and lower bounds as we will see on the next slide) for probability statements and not the exact probabilities.

Other forms of Chebyshev Inequality

Recall that $\mathbb{V}(X) = \sigma^2$ and $\sqrt{\mathbb{V}(X)} = \sigma$. If we let $t = a\sigma$ for some $a > 0$, then Chebyshev's Inequality has a slightly different form:

$$Pr\left(|X - E(X)| \geq a\sigma\right) \leq \frac{1}{a^2}$$

This gives us an idea of the proportion of data that lie outside "a" standard deviations of the mean.

It also gives an idea of the probability that any given random variable will differ from its mean. That is, the probability that X will differ from $E(X)$ by more than "a" standard deviations cannot exceed $1/a^2$.

Other forms of Chebyshev Inequality

If we would like to learn about the proportion of data that lie within "a" standard deviations of the mean, then we can adjust the inequality yet again to have:

$$Pr\left(|X - E(X)| < a\sigma\right) \geq 1 - \frac{1}{a^2}$$

Then the followings are true:

- At least 75% of the observations must always be less than 2 standard deviations from the population mean.
- At least 89% of the observations must always be less than 3 standard deviations of the population mean.

What proportion of the observations must lie wthin 4 standard deviations of the population mean?

Law of Large Numbers (LLN)

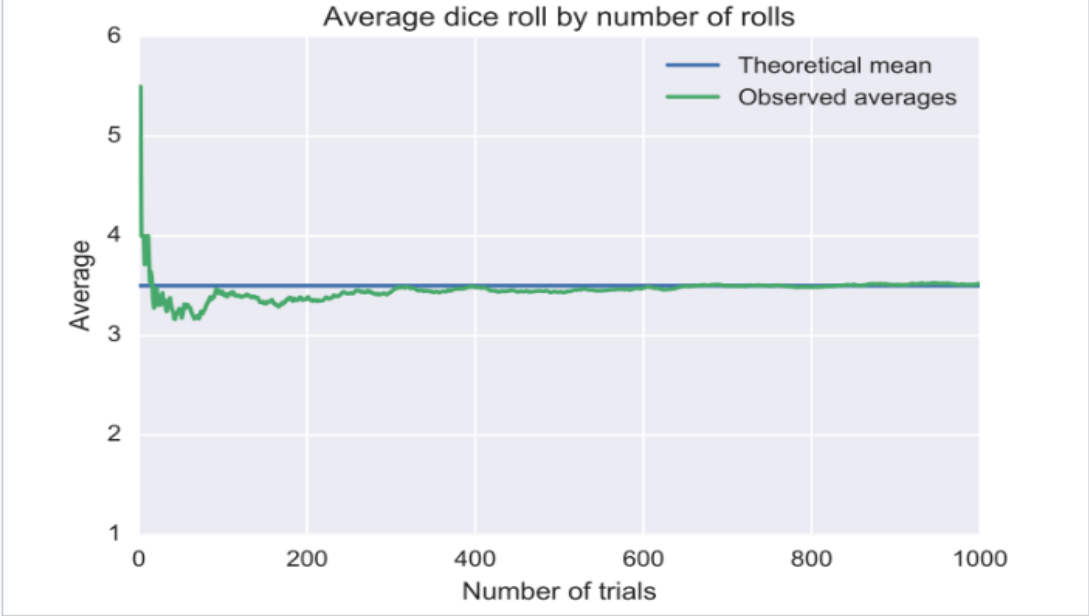
Let $X_1, X_2, ..., X_n$ be independent and identically distributed (i.i.d.) random variables with finite expectation $\mu = \mathbb{E}(X_i)$, and let the sample mean be $\overline{X_n} = \sum_{i=1}^{n} X_i/n$. Then

$$P(\overline{X_n} \to \mu) = 1, \qquad \text{as} \quad n \to \infty$$

In words, the **Law of Large Numbers (LLN)** shows that sample averages converge to the population mean μ (with probability 1) as the sample size increases.

– As more observations are collected, the proportion of occurrences with a particular outcome, $\widehat{p_n}$, converges to the probability of that outcome p.

An Illustration



An illustration of the law of large numbers using a particular run of rolls of a single die. As the number of rolls in this run increases, the average of the values of all the results approaches 3.5. While different runs would show a different shape over a small number of throws (at the left), over a large number of rolls (to the right) they would be extremely similar.

https://en.wikipedia.org/wiki/Law_of_large_numbers

The sample mean stabilizes to the expected value 3.5 as n increases.

Random Sample

A **random sample** is a sample generated by making repeated draws from a box containing numbers or in the case of any random variable X, from the distribution of X. Clearly, for continuous distributions, there will be an infinite number of options/values of X.

In a random sample, draws are made **with replacement**, and the outcomes in a series of draws are independent. The value on the first draw does not affect the value on the second.

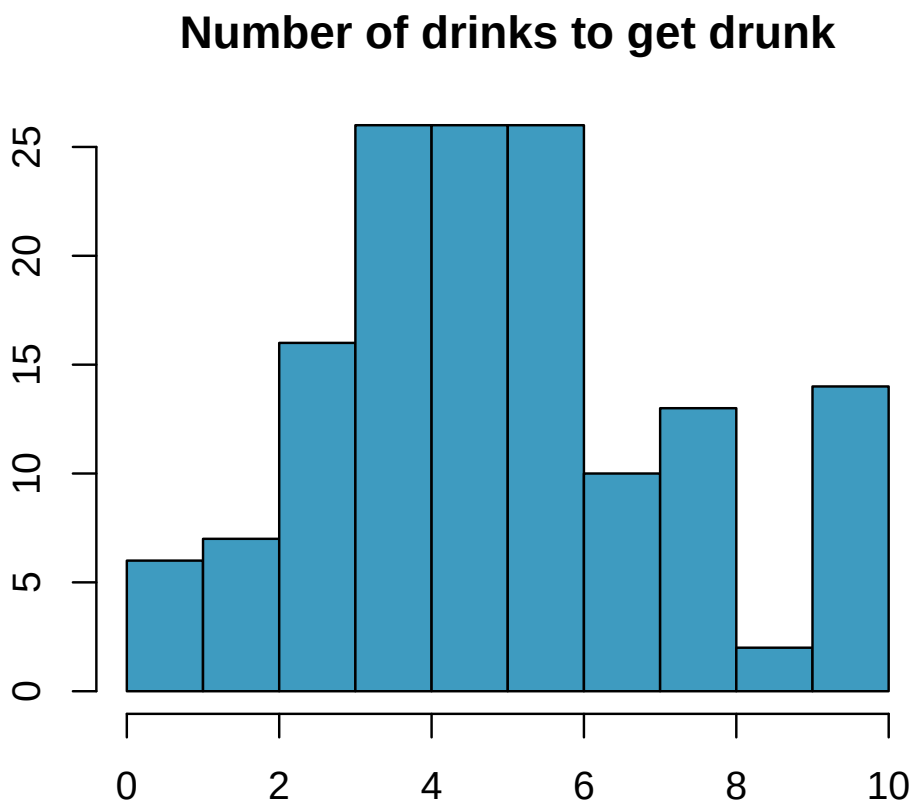
Population v.s. Sample

- We are often interested in **population characteristics**.
- Since complete populations are difficult (or impossible) to collect data on, we use **sample statistics** as an estimate for the unknown population characteristics of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample statistics vary provides a way to estimate the **margin of error** associated with our estimate.
- Let's try to understand how and why sample statistics varies from sample to sample.

Example: suppose we randomly sample 1,000 adults **from each state** in the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Not the same, but only somewhat different.

The following histogram shows the distribution of number of drinks it takes a group of college students to get drunk. We will assume that this is our population of interest. If we randomly select observations from this data set, which values are most likely to be selected, which are least likely?



Suppose that you don't have access to the population data. In order to estimate the average number of drinks it takes these college students to get drunk, you might sample from the population and use your sample mean as the best guess for the unknown population mean.

- Sample, with replacement, ten students from the population, and record the number of drinks it takes them to get drunk.
- Find the sample mean.
- Plot the distribution of the sample averages obtained by sampling multiple times.

1	7	16	3	31	5	46	4	61	10	76	6	91	4	106	6	121	6	136	6
2	5	17	10	32	9	47	3	62	7	77	6	92	0.5	107	2	122	5	137	7
3	4	18	8	33	7	48	3	63	4	78	5	93	3	108	5	123	3	138	3
4	4	19	5	34	5	49	6	64	5	79	4	94	3	109	1	124	2	139	10
5	6	20	10	35	5	50	8	65	6	80	5	95	5	110	5	125	2	140	4
6	2	21	6	36	7	51	8	66	6	81	6	96	6	111	5	126	5	141	4
7	3	22	2	37	4	52	8	67	6	82	5	97	4	112	4	127	10	142	6
8	5	23	6	38	0	53	2	68	7	83	6	98	4	113	4	128	4	143	6
9	5	24	7	39	4	54	4	69	7	84	8	99	2	114	9	129	1	144	4
10	6	25	3	40	3	55	8	70	5	85	4	100	5	115	4	130	4	145	5
11	1	26	6	41	6	56	3	71	10	86	10	101	4	116	3	131	10	146	5
12	10	27	5	42	10	57	5	72	3	87	5	102	7	117	3	132	8		
13	4	28	8	43	3	58	5	73	5.5	88	10	103	6	118	4	133	10		
14	4	29	0	44	6	59	8	74	7	89	8	104	8	119	4	134	6		
15	6	30	8	45	10	60	4	75	10	90	5	105	3	120	8	135	6		

Example:

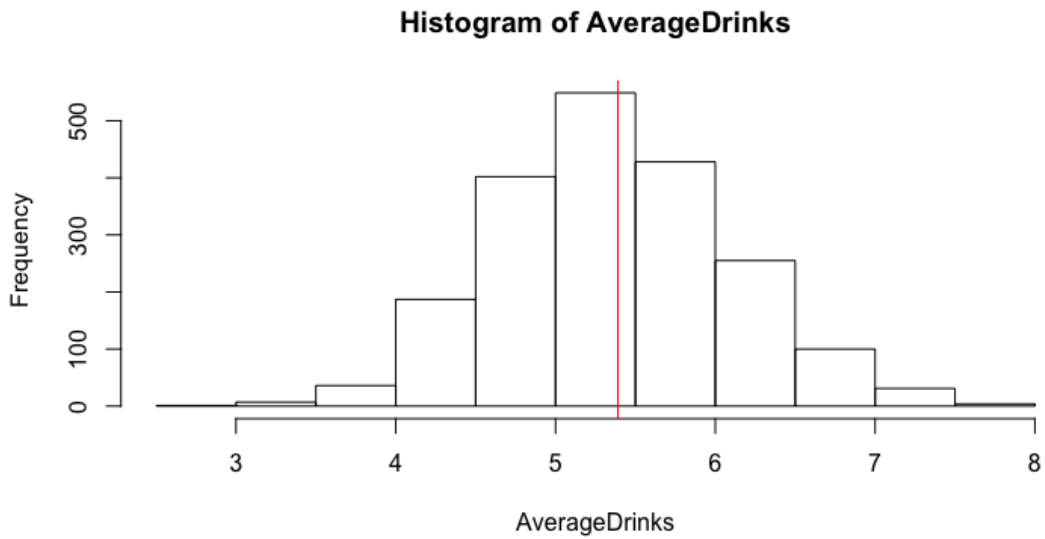
List of random numbers: 59, 121, 88, 46, 58, 72, 82, 81, 5, 10

1	7	16	3	31	5	46	4	61	10	76	6	91	4	106	6	121	6	136	6
2	5	17	10	32	9	47	3	62	7	77	6	92	0.5	107	2	122	5	137	7
3	4	18	8	33	7	48	3	63	4	78	5	93	3	108	5	123	3	138	3
4	4	19	5	34	5	49	6	64	5	79	4	94	3	109	1	124	2	139	10
5	6	20	10	35	5	50	8	65	6	80	5	95	5	110	5	125	2	140	4
6	2	21	6	36	7	51	8	66	6	81	6	96	6	111	5	126	5	141	4
7	3	22	2	37	4	52	8	67	6	82	5	97	4	112	4	127	10	142	6
8	5	23	6	38	0	53	2	68	7	83	6	98	4	113	4	128	4	143	6
9	5	24	7	39	4	54	4	69	7	84	8	99	2	114	9	129	1	144	4
10	6	25	3	40	3	55	8	70	5	85	4	100	5	115	4	130	4	145	5
11	1	26	6	41	6	56	3	71	10	86	10	101	4	116	3	131	10	146	5
12	10	27	5	42	10	57	5	72	3	87	5	102	7	117	3	132	8		
13	4	28	8	43	3	58	5	73	5.5	88	10	103	6	118	4	133	10		
14	4	29	0	44	6	59	8	74	7	89	8	104	8	119	4	134	6		
15	6	30	8	45	10	60	4	75	10	90	5	105	3	120	8	135	6		

Sample mean: (8+6+10+4+5+3+5+6+6+6) / 10 = 5.9

Repeat the process multiple times

Sampling Distribution



What we just constructed is called a **sampling distribution**.

What is the shape and center of this distribution? Based on this distribution, what do you think is the true population average?

Approximately 5.39, the true population mean.

Is the symmetric bell shape a coincidence?

Sums and Averages of Random Variables

Assume that $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with finite mean and variance $\mu = E(X_i)$ and $\sigma^2 = V(X)$, and let $S_n = X_1 + X_2 + \ldots + X_n$ and $\overline{X_n} = S_n/n$. Using properties of expectations and variances,

$$E(S_n) = E(X_1) + E(X_2) + \ldots + E(X_n) = n\mu$$
$$V(S_n) = V(X_1) + V(X_2) + \ldots + V(X_n) = n\sigma^2$$

and similarly for $\overline{X_n}$:

$$E(\overline{X_n}) = \frac{1}{n}\Big(E(X_1) + E(X_2) + \ldots + E(X_n)\Big) = \frac{1}{n}\Big(\mu + \mu + \ldots + \mu\Big) = \mu$$
$$V(\overline{X_n}) = \frac{1}{n^2}\Big(V(X_1) + V(X_2) + \ldots + V(X_n)\Big) = \frac{1}{n^2}\Big(\sigma^2 + \sigma^2 + \ldots + \sigma^2\Big) = \frac{\sigma^2}{n}$$

The variance of the sample mean $V(\overline{X_n})$ shrinks to zero as $n \to \infty$.

Sums and Averages of Normals

Now suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. Normal(μ, σ^2). Since linear combinations of Normals are Normal:

$$\overline{X_n} \sim \text{Normal}(\mu, \sigma^2/n), \quad S_n \sim \text{Normal}(n\mu, n\sigma^2)$$

so we can standardize and compute probabilities using the standard Normal table, that is

$$P\left(\frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma} \leq z\right) = P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z\right) = P(Z \leq z)$$

where $Z \sim \text{Normal}(0, 1)$.

Central Limit Theorem (CLT)

If the sample size is big enough, we can do the same thing with sums and averages of random variables that **are not necessarily Normal**. More precisely, let $X_1, X_2, \ldots, X_n$ be i.i.d. with finite expectation and variance $E(X_i) = \mu$ and $V(X_i) = \sigma^2$. Let $S_n = X_1 + X_2 + \ldots + X_n$ and $\overline{X_n} = S_n/n$, and $Z \sim \text{Normal}(0, 1)$. Then

$$P\left(\frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma} \leq z\right) = P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z\right) \to P(Z \leq z)$$

as $n \to \infty$.

In practice, we will use CLT to approximate the distributions of sums and averages of random variable by Normal distributions,

$$\overline{X_n} \overset{d}{\approx} \text{Normal}(\mu, \sigma^2/n), \quad S_n \overset{d}{\approx} \text{Normal}(n\mu, n\sigma^2)$$

History of Central Limit Theorem

The **Central Limit Theorem (CLT)** is one of the high-water marks of mathematical thinking. It was worked upon by James Bernoulli, Abraham de Moivre, and Alan Turing. Over the centuries, the theory improved from special cases to a very general rule.

Essentially, the Central Limit Theorem allows one to describe how accurately the Law of Averages works. Most people have a good intuitive understanding of the Law of Averages, but in many cases it is important to determine whether a particular size of deviation between the sample mean and the (usually unknown) expected value is probable or improbable. That is, what is the chance that the sample average is more than some constant "a" away from the true μ?

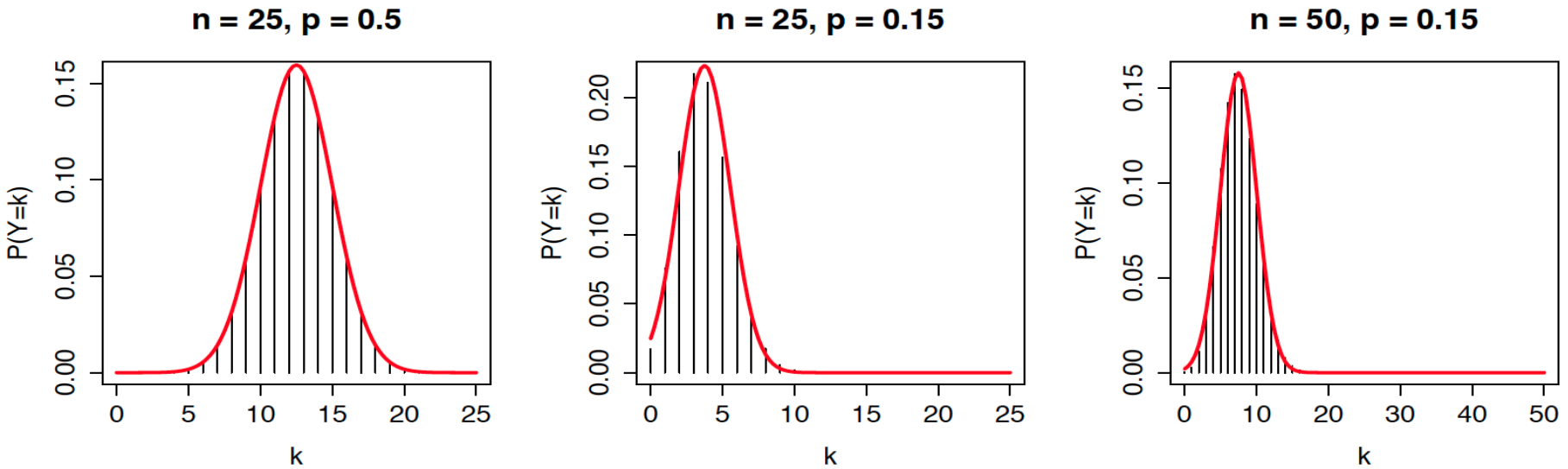
Conditions for Central Limit Theorem

Certain conditions must be met for the CLT to apply:

1. **Independence:** Sampled observations must be independent. This is difficult to verify, but is more likely if
   - Random sampling/assignment is used, and
   - If sampling without replacement, $n < 10\%$ of the population
2. **Sample size:** Either the population distribution is normal, or if the population distribution is skewed,
   - the more skewed the population distribution is, the larger sample size we need for the CLT to apply
   - for moderately skewed distributions, $n > 30$ is a widely used rule of thumb

Examples: Binomial

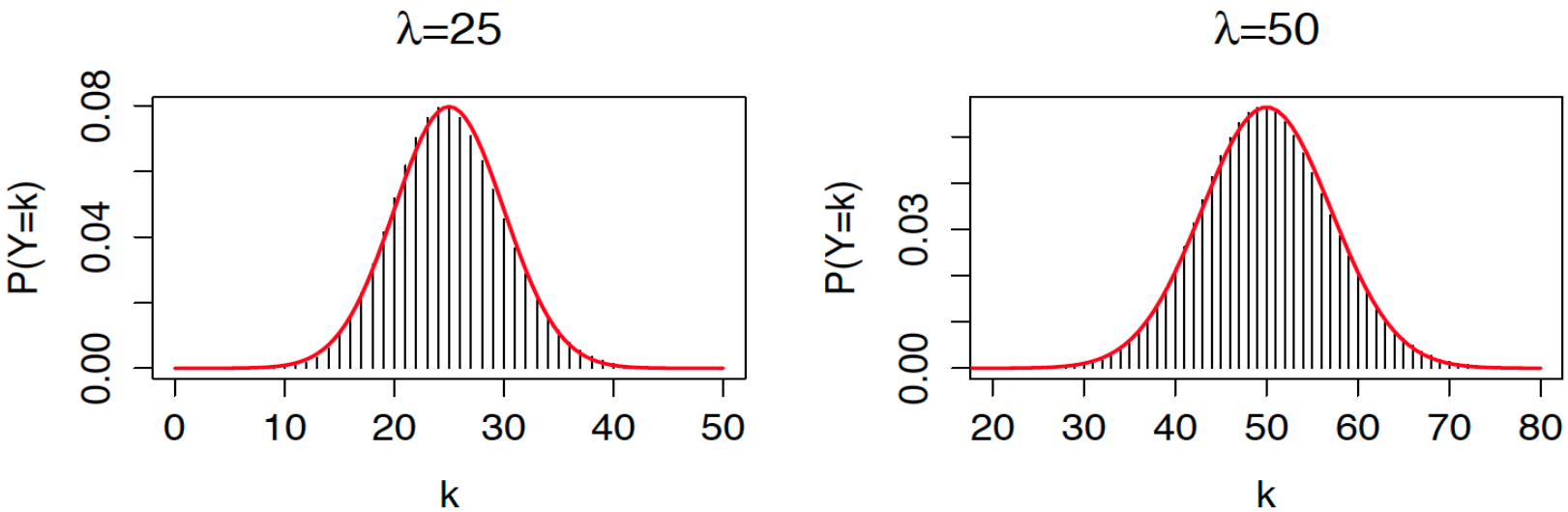
If $Y \sim \text{Binomial}(n, p)$, we can write $Y = X_1 + X_2 + \ldots + X_n$, where X_i are independent Bernoulli(p) random variables. By CLT, we know that $Y \overset{d}{\approx} \text{Normal}(np, np(1-p))$. The figures below show the Binomial PMF and the pdf of the Normal approximation for $(n, p) \in \{(25, 0.5), (25, 0.15), (50, 0.15)\}$.



The Binomial distribution is symmetric if $p = 0.5$ and it becomes more skewed as we approach extreme values (close to 0 or 1).

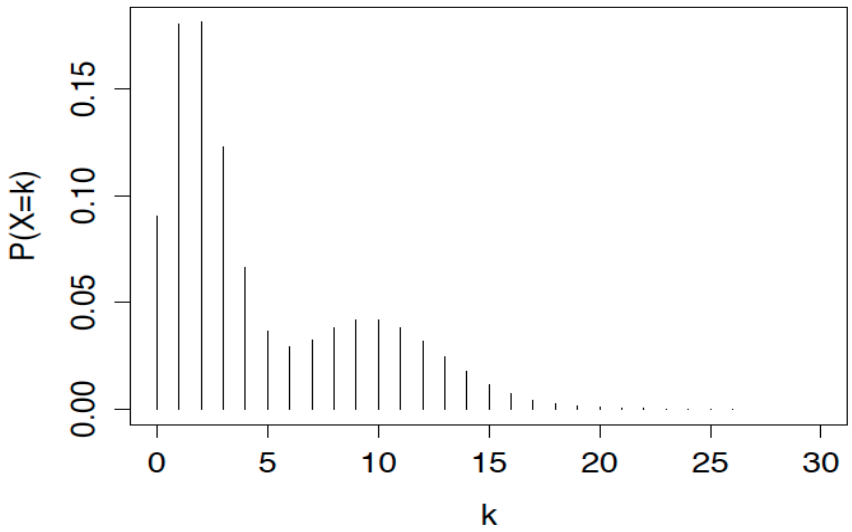
Examples: Poisson

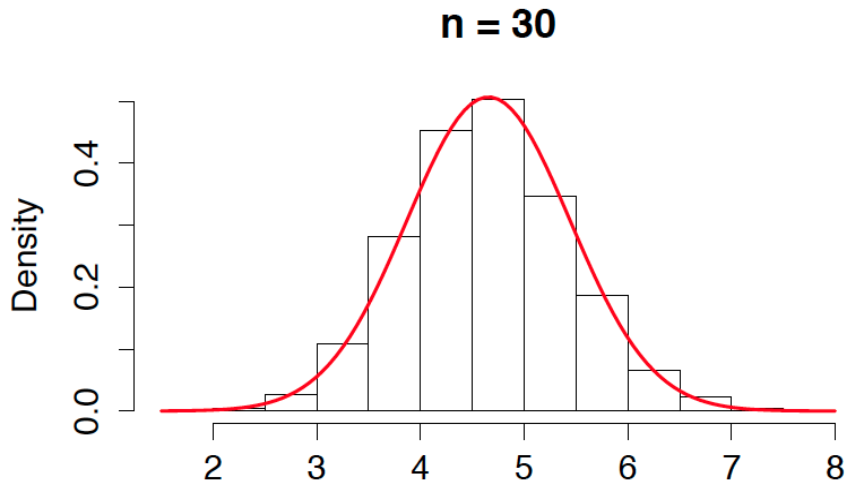
If $Y \sim \text{Poisson}(\lambda)$, we can write $Y = X_1 + X_2 + \ldots + X_n$, where X_i are independent $\text{Poisson}(\lambda/n)$ random variables. By CLT, we can approximate the Poisson as a $\text{Normal}(\lambda, \lambda)$. The figures below show the Poisson PMF for $\lambda \in \{25, 50\}$ and the pdf of the corresponding Normal approximation



Examples: a Weird Distribution

Let $X_1, X_2, \ldots, X_n$ be i.i.d drawn from a discrete distribution with the following PMF (left panel):





The figure in the right panel shows a histogram found after taking 5000 averages from samples of size $n = 30$ coming from the weird distribution (and the corresponding Normal approximation)

This is telling us that the distribution of sample averages coming from this odd-looking discrete distribution can be approximated well with a Normal distribution, even if n is as small as 30.

Examples: Sum of Random Variables

Problem 1: You are playing Red and Black in roulette. (A roulette wheel has 38 pockets; 18 are red, 18 are black, and 2 are green – the house takes all the money on green). Let's assume equal probability for each pocket.

You pick either red or black; if the ball lands in the color you pick, you win a dollar. Otherwise you lose a dollar.

Suppose you make 100 plays. What is the chance that you lose $10 or more?

Every time you make a play, there are 38 tickets, and 18 are labelled 1 and the 20 are labelled -1.

Examples: Sum of Random Variables (Cont'd)

So the theoretical expected value is

$$\begin{aligned}\mu &= \frac{1}{38}[1+1+\cdots+1+(-1)+(-1)+\cdots+(-1)] \\ &= \frac{1}{38}[-2] \\ &= -1/19.\end{aligned}$$

The theoretical standard deviation is

$$\begin{aligned}\sigma &= \sqrt{(\frac{1}{38}\sum_{i=1}^{38} X_i^2) - \mu^2} \\ &= \sqrt{1-(-1/19)^2} \\ &= .998614.\end{aligned}$$

Examples: Sum of Random Variables (Cont'd)

The probability of losing more than \$10 or more in 100 plays is

$$\begin{aligned}
\mathrm{P}\left[\sum_i^{100} Y_i < -10\right] &= \mathrm{P}\left[\sum_i^{100} Y_i - n\mu < -10 - n\mu\right] \\
&= \mathrm{P}\left[\frac{\sum_i^{100} Y_i - n\mu}{\sigma\sqrt{n}} < \frac{-10 - n\mu}{\sigma\sqrt{n}}\right] \\
&\doteq \mathrm{P}\left[Z < \frac{-10 - n\mu}{\sigma\sqrt{n}}\right] \\
&= \mathrm{P}\left[Z < \frac{[-10 - (100)(-1/19)]}{(10 * .998614)}\right] \\
&= \mathrm{P}[Z < -0.47434].
\end{aligned}$$

From the standard normal table, the chance of being below -0.47434 is about 0.319.

Examples: Average of Random Variables

Problem 2a: You want to estimate the average income of people in Durham. Suppose the true mean income μ is \$42,000 with standard deviation σ of \$10,000. You draw a random sample of 100 households.

What is the **probability** that your **sample mean** is over the **true average income** by \$500 or more?

Note that in order to solve this, we made the unreasonable assumption that we knew the true mean and the standard deviation. Later we shall relax this assumption.

Examples: Average of Random Variables (Cont'd)

Since we know μ and σ, this is relatively straightforward.

$$\begin{aligned}
\mathbb{P}[\overline{X} > 42,500] &= \mathbb{P}[\overline{X} - \mu > 500] \\
&= \mathbb{P}\left[\frac{\overline{X} - \mu}{(\sigma/\sqrt{n})} > \frac{500}{(\sigma/\sqrt{n})}\right] \\
&\doteq \mathbb{P}\left[Z > \frac{500}{(10,000/\sqrt{100})}\right] \\
&= \mathbb{P}[Z > .5].
\end{aligned}$$

The CLT is used in the penultimate step.

From the standard normal table, we know this has chance 0.309. There is about a 31% chance of $\overline{X}$ being too high by \$500 or more.

But this is not really the question one wants to ask in practice, nor is it the kind of information that one really has from a survey.

Examples: Average of Random Variables (Cont'd)

Problem 2b: You want to estimate the average income of people in Durham. You draw a random sample of 100 households and find the sample mean $\overline{X}$ is \$42,500 and the standard deviation SD of your sample is \$10,000.

What is the **approximate probability** that you have overestimated the **true average income** by \$500 or more?

First, assume that the SD of your sample equals the true standard deviation σ. This is an approximation, and later we shall see a way to improve this. **Despite that μ is unknown**, we are still able to find

$$\begin{aligned}
\mathbb{P}[\overline{X} - \mu > 500] &= \mathbb{P}\left[\frac{\overline{X} - \mu}{(\sigma/\sqrt{n})} > \frac{500}{(\sigma/\sqrt{n})}\right] \\
&\doteq \mathbb{P}\left[Z > \frac{500}{(10,000/\sqrt{100})}\right] \\
&= \mathbb{P}[Z > .5].
\end{aligned}$$

The probability is 0.309 that \$42,500 is \$500 (or more) too high.

Finite Population Correction Factor

The CLT and the standard errors (deviations) of sample averages or mean are based on samples selected **with replacement**. However, in virtually all survey research, you sample **without replacement** from populations that are of a finite size, M.

If the sample size n is small compared to the population size M, one can ignore the distinction between sampling with replacement and without replacement. Then the standard deviation of an average is $\sigma / \sqrt{n}$.

If n is large relative to the population size M (say n, is more than 5% of the population size, M), then use the **Finite Population Correction Factor (FPCF)** to find the standard deviation of the average as

$$\frac{\sigma}{\sqrt{n}} * \sqrt{\frac{M-n}{M-1}}.$$

Recap

Today we covered:

- Linear combination of random variables
- Markov and Chebyshev Inequalities
- Law of large numbers (LLN)
- Variability in Estimates
- The central limit theorem (CLT)

Suggested reading:

- D.S. Sec. 6.1, 6.2, 6.3
- OpenIntro3: Sec. 4.4