Week 2

Variance

The population variance

The variance of a random variable is a measure of spread.

$$\operatorname{Var}(X) = \mathrm{E}[(X - \mu)^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$$

where $\mathrm{E}[X^2] = \sum_{x \in \Omega} x^2 p(x)$, for discrete variable, or $\mathrm{E}[X^2] = \int_{\Omega} x^2 f(x) dx$, for continuous variables.

The square root of the variance is called **standard deviation**.

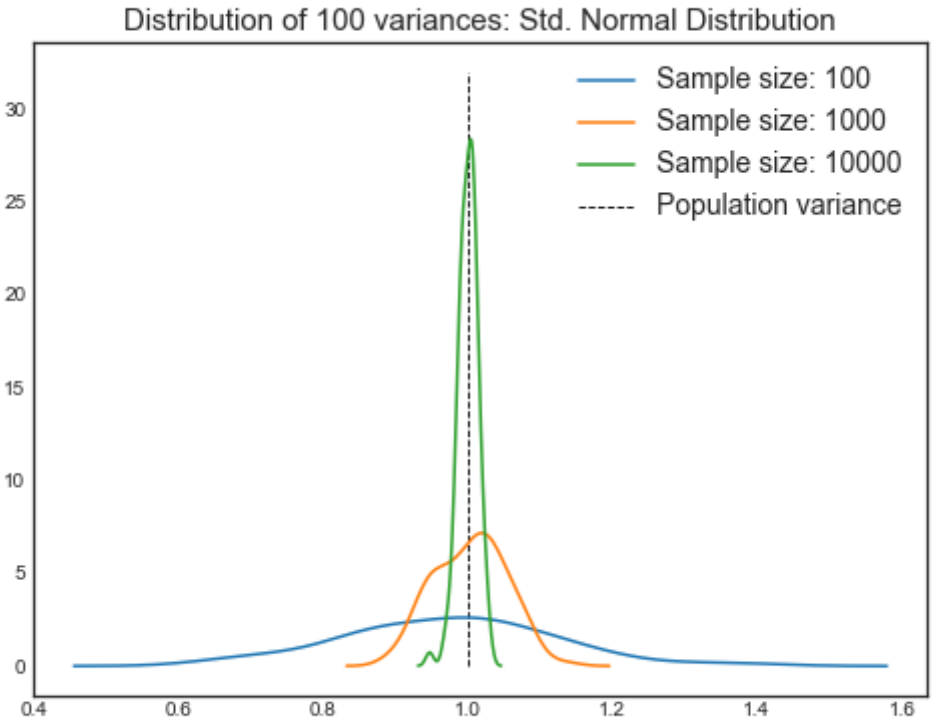
The sample variance

The sample variance is the average squared distance of each observed data point to the sample mean.

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

The square root of the sample variance is the sample standard deviation.

- The sample variance is a random variable;
- As such, it has an associate population distribution;
- Whose expected value is the population variance.
- More data will produce a more concentrated distribution around the expected value (population variance).



Recall the mean

Recall that the average of a random sample from a population is itself a random variable.

$$\mathrm{E}[\bar{X}] = \mu$$

where μ is the population mean.

So, the variance of the sample mean decreases to zero as it accumulates more data:

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

where σ^2 is the population variance and n, the sample size.

We cal the standard deviation of a statistic a standard error.

- So, the square root of the variation of the mean is the standard error of the mean.

Since the variance of the sample mean is $\frac{\sigma^2}{n}$:

- Its logical estimate is $\frac{s^2}{n}$;
- The logical estimate of the standard error is $\frac{S}{\sqrt{n}}$.
- The standard error talks about how variable averages of random samples of size n from the population are.

Facts about the variance

- $$\begin{aligned}\mathrm{Var}(aX) &= \mathrm{E}[(aX)^2] - \big(\mathrm{E}[aX]\big)^2 \\ &= \mathrm{E}[a^2X^2] - \big(a\mathrm{E}[X]\big)^2 \\ &= a^2\mathrm{E}[X^2] - a^2\big(\mathrm{E}[X]\big)^2 \\ &= a^2\mathrm{Var}(X)\end{aligned}$$

- The sample variance estimates the population variance;
- The distribution of the sample variance is centered at what it is estimating (the population variance);
    - This means that the sample variance is unbiased;
    - The distribution gets more concentrated around the population variance with larger sample sizes;
- The variance of the sample mean is the population mean divided by the sample size: $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$.
    - Its square root $\frac{\sigma}{\sqrt{n}}$ is the standard error of the mean.

Distributions

Bernoulli distribution

The Bernoulli distribution arises as the result of a binary outcome.

Bernoulli random variables take only the values 1 and 0, with probabilities p and $1-p$, respectively.

PMF:

$$\mathrm{P}(X = x) = p^x (1-p)^{1-x}$$

- The mean of a Bernoulli random variable is p ;
- The variance is $p(1-p)$.

Binomial distribution

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\mathrm{Bernoulli}(p)$.

Then, $X = \sum_{i=1}^{n} X_i$ is a binomial random variable.

PMF:

$$\mathrm{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$.

- The mean of a binomial random variable is np ;
- The variance is $np(1-p)$.

Normal distribution

A random variable is said to follow a normal (or Gaussian) distribution ($X \sim \mathcal{N}(\mu, \sigma^2)$) with mean μ and variance σ^2 if its PDF is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

with $\mathrm{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

- The distribution for which $\mu = 1$ and $\sigma = 1$ is known as standard normal distribution (often labeled z).
- The area under the curve between $\mu - \sigma$ and $\mu + \sigma$, that is, the probability $\mathrm{P}\big(\mu - \sigma \leq X \leq \mu + \sigma\big)$ is approximately 0.6827 (or 68.27%).
- The probability $\mathrm{P}\big(\mu - 2\sigma \leq X \leq \mu + 2\sigma\big)$ is approximately 0.9545 (or 95.45%).
- The probability $\mathrm{P}\big(\mu - 3\sigma \leq X \leq \mu + 3\sigma\big)$ is approximately 0.9973 (or 99.73%).

Facts about the normal distribution:

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \dfrac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$.
- If Z is a standard normal variable, then $X = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$.
- The values -1.28, -1.645, -1.96 and -2.33 are, respectively, the 10^{th}, 5^{th}, 2.5^{th}, and 1^{st} percentiles of the standard normal distribution.
- By symmetry, 1.28, 1.645, 1.96 and 2.33 are, respectively, the 90^{th}, 95^{th}, 97.5^{th}, and 99^{th} percentiles of the standard normal distribution.

Question:

What is the 95^{th} percentile of a $\mathcal{N}(\mu, \sigma^2)$ distribution?

$$z_p = \frac{x_p - \mu}{\sigma},$$

where z_p is the p^{th} percentile of the standard normal distribution.

- For the 95^{th} percentile, $z_p = 1.96$, as seen above.
- Then, x_p, the 95^{th} percentile of the normal distribution of mean μ and variance σ^2, is given by: $x_p = \mu + 1.96\sigma$.

Poisson distribution

Used to model counts.

PMF:

$$\mathrm{P}(X = x; \lambda) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!}$$

with $x \in \mathbb{N}$.

- The mean of this distribution is λ.
- The variance, too, is λ.

Some uses for the Poisson distribution:

- Modeling count data;
- Modeling event-time or survival data;
- Modeling contingency tables;
- Approximating binomials when n is large and p is small.

Rates and Poisson random variables:

- Poisson random variables are used to model rates;
- $X \sim \mathrm{Poisson}(\lambda t)$ where:
  - $\lambda = \mathrm{E}\left[\dfrac{X}{t}\right]$ is the expected count per unit of time;
  - t is the total monitoring time.

Poisson approximation to the binomial:

- Let $X \sim \text{Binomial}(n, p)$.
- For $n \gg 1$ and $p \ll 1$, X can be approximated by a Poisson distribution, with $\lambda = np$.

Asymptotics

Asymptotics is the term for the behavior of statistics as the sample size (or some other quantity) limits to infinity (or some other relevant number).

Limits of random variables

These results allows us to talk about the large sample distribution of sample means of a collection of i.i.d. observations.

The first of these results, the Law of Large Numbers (LLN), we intuitively know:

- It says that the average limits to what it is estimating, the population mean.
  - For instance, let $\bar{X}_n$ be the average of the result of n coin flips.
  - As we flip a coin over and over, it eventually converges to the expected value, that is, $\bar{X}_n \to \mu \because n \to \infty$.
- According to the LLN, the sample mean of i.i.d. samples is a consistent estimator of the population mean.
  - The sample variance and the sample standard deviation are consistent, as well.

> An estimator is said to be consistent if it converges, as the sample size grows, to the estimand.

The Central Limit Theorem

For our purposes, the CLT states that the distribution of averages of i.i.d. variables (properly normalized) becomes that of a standard normal, as the sample size increases.

The idea is that, for an estimate, subtracting the mean of that estimate and dividing by the standard error of said estimate, one obtains a random variable whose distribution increasingly approximates a standard normal distribution.

$$\frac{\bar{X}_n - \text{E}[\bar{X}_n]}{\sqrt{\text{Var}(\bar{X}_n)}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1), \quad n \to \infty$$

In other words, the sample mean approximately follows a normal distribution of mean μ and variance $\frac{\sigma^2}{n}$.

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Example:

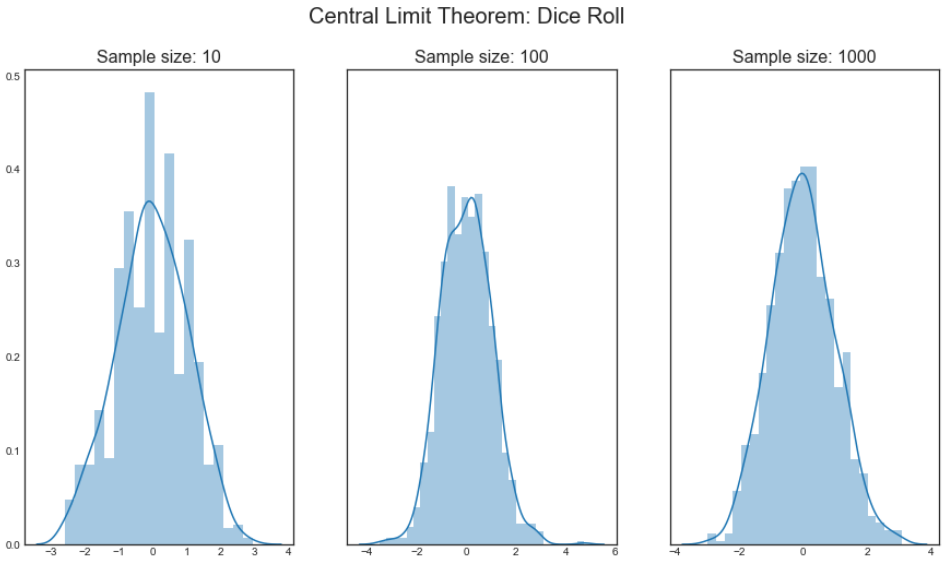
Let X_i be the outcome for die i.

Note that $\mu = \text{E}[X_i] = 3.5$ and $\text{Var}(X_i) = 2.92$.

Then, the standard error of the mean is $\text{SE} = \sqrt{\frac{2.92}{n}} \approx \frac{1.71}{\sqrt{n}}$.

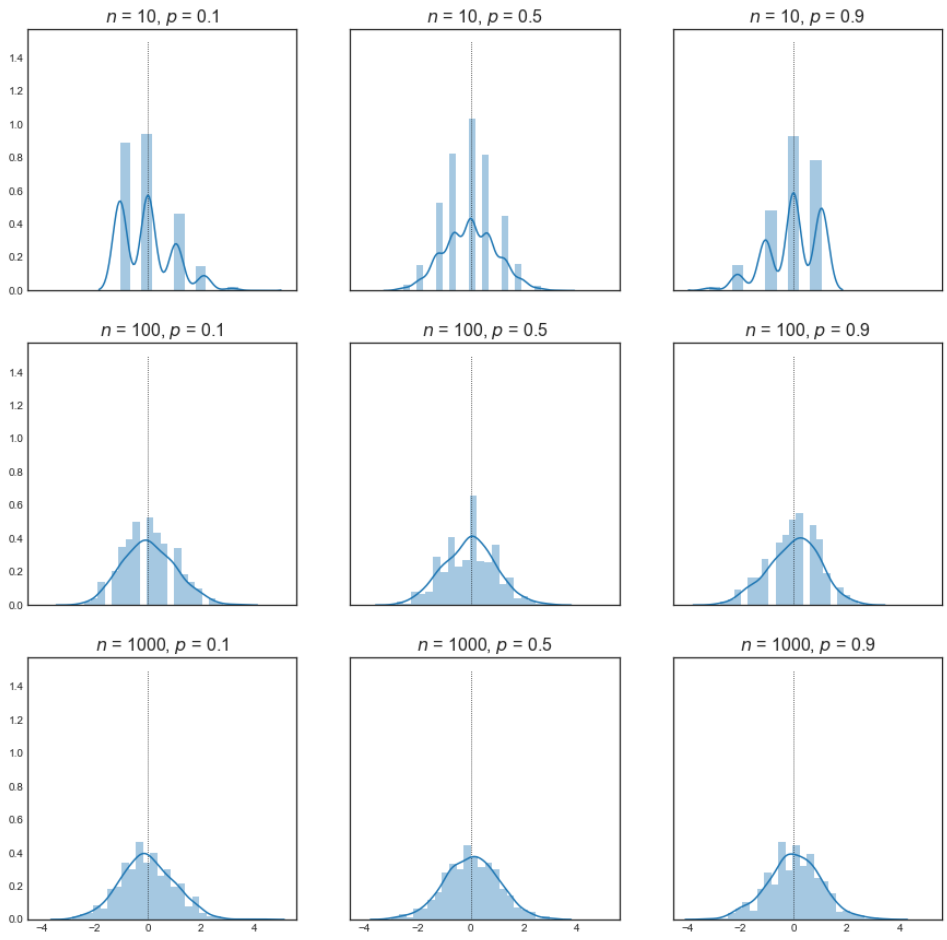
Let us roll n dice, take their mean, subtract off 3.5 and divide by SE.

If the CLT is right, the result will resemble a bell curve.



Now, let X_i be the 0 or 1 result of the i^{th} flip of a possibly unfair coin.

- The sample proportion, say $\hat{p}$, is the average of the coin flips;
- $\mathrm{E}[X_i] = p$ and $\mathrm{Var}(X_i) = p(1-p)$;
- Standard error of the mean is $\sqrt{\frac{p(1-p)}{n}}$;
- Then, $\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0,1)$.



> "The speed at which the normalized coin flips converge to normality is governed by how bias the coin is" (the difference between p and $1-p$).

Confidence intervals

$\bar{X}$ is approximately normal with mean μ and std. deviation $\frac{\sigma}{\sqrt{n}}$.

- Then, the probability that $\bar{X}$ is bigger than $\mu + 2\frac{\sigma}{\sqrt{n}}$ or smaller than $\mu - 2\frac{\sigma}{\sqrt{n}}$ is (roughly) 5%: $\mathrm{P}\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.9545.$
- $\bar{X} \pm 2\frac{\sigma}{\sqrt{n}}$ is called 95% confidence interval for μ.

> "The actual interpretation of this is that, if we were to repeatedly get samples of size n from this population, construct a confidence interval in each case, about 95% would contain μ".

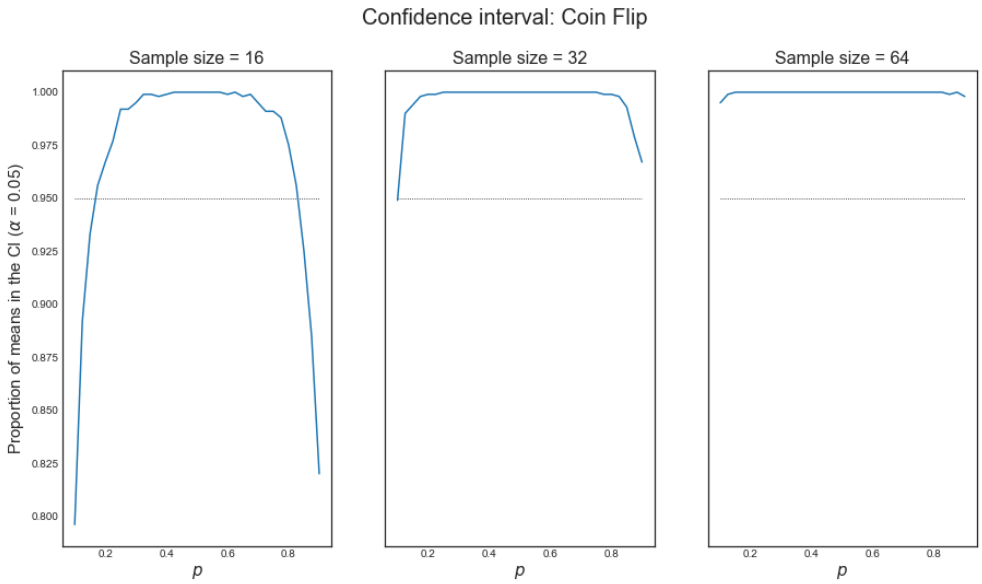
More generally, for a level of significance α (that is, confidence $1-\alpha$), the confidence interval is given by:

$$\mathrm{CI}_{1-\alpha} = \bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)^{\text{th}}$ percentile of a standard normal distribution.

Sample proportions

- In the event that each X_i is 0 or 1 with common success probability p, then $\sigma^2 = p(1-p)$;
- The interval takes the form $\hat{p} = z_{1-\alpha/2}\sqrt{\dfrac{p(1-p)}{n}}$.
    - Replacing p by $\hat{p}$ in the standard error results in what is called a Wald confidence interval for p.
    - For 95% intervals, $\hat{p} \pm \dfrac{1}{\sqrt{n}}$ is a quick CI estimate for p.



Tip:

- If n is not large enough for the CLT to be applicable for many of the values of p:
    - Quick fix, form the interval with $\hat{p} = \dfrac{X+2}{n+4}$.
    - (Add two successes and failures: Agresti-Coull interval).

Poisson interval

A nuclear pump failed 5 times out of 94.32 days, give a 95% confidence interval for the failure rate per day.

- $X \sim \mathrm{Poisson}(\lambda t)$;
- Estimate $\hat{\lambda} = \dfrac{X}{t}$;
- $\mathrm{Var}(\hat{\lambda}) = \dfrac{\lambda}{t}$;
- $\dfrac{\hat{\lambda}}{t}$ is our variance estimate.
- Then, $\mathrm{CI}_{1-\alpha} = \hat{\lambda} \pm z_{1-\alpha/2}\sqrt{\dfrac{\hat{\lambda}}{t}}$.