Sampling

DSE 210

Review: expected value

The expected value of a random variable X is

$$\mathbb{E}(X) = \sum_x x \Pr(X = x).$$

Example: A coin has heads probability p. Let X be 1 if heads, 0 if tails.

$$\mathbb{E}(X) \;=\; 1 \cdot p + 0 \cdot (1 - p) \;=\; p.$$

Linearity properties:

- $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$ for any random variable X and any constants a, b.
- $\mathbb{E}(X_1 + \cdots + X_k) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_k)$ for any random variables $X_1, X_2, \ldots, X_k$.

Example: Toss n coins of bias p, and let X be the number of heads. What is $\mathbb{E}(X)$?

Let the individual coins be $X_1, \ldots, X_n$.

$$\mathbb{E}(X) = \mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) = np.$$

Review: variance

$\text{var}(X) = \mathbb{E}(X-\mu)^2 = \mathbb{E}(X^2) - \mu^2$, where $\mu = \mathbb{E}(X)$.

Toss a coin of bias p. Let $X \in \{0, 1\}$ be the outcome.

$$\mathbb{E}(X) = p$$
$$\mathbb{E}(X^2) = p$$
$$\mathbb{E}(X-\mu)^2 = p^2 \cdot (1-p) + (1-p)^2 \cdot p = p(1-p)$$
$$\mathbb{E}(X^2) - \mu^2 = p - p^2 = p(1-p)$$

This variance is highest when $p = 1/2$ (fair coin).

The standard deviation of X is $\sqrt{\text{var}(X)}$.
It is the average amount by which X differs from its mean.

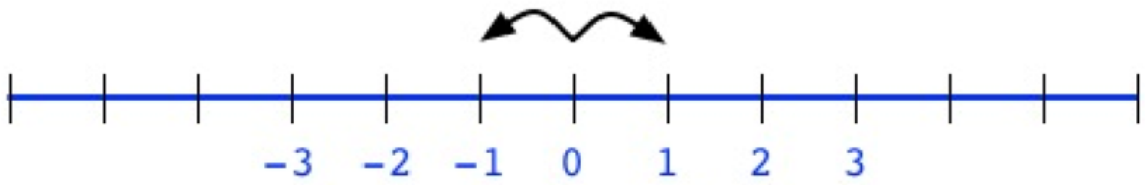
Useful variance rules:

- $\text{var}(X_1 + \cdots + X_k) = \text{var}(X_1) + \cdots + \text{var}(X_k)$ if X_i's independent.
- $\text{var}(aX + b) = a^2\text{var}(X)$.

Variance of a sum

$\text{var}(X_1 + \cdots + X_k) = \text{var}(X_1) + \cdots + \text{var}(X_k)$ if the X_i are independent.

Symmetric random walk. A drunken man sets out from a bar. At each time step, he either moves one step to the right or one step to the left, with equal probabilities. Roughly where is he after n steps?



Let $X_i \in \{-1, 1\}$ be his ith step. Then $\mathbb{E}(X_i) = 0$ and $\text{var}(X_i) = 1$.

His position after n steps is $X = X_1 + \cdots + X_n$.

$$\mathbb{E}(X) = 0$$
$$\text{var}(X) = n$$
$$\text{stddev}(X) = \sqrt{n}$$

What is the distribution over his possible positions?
Approximately $N(0, n)$: Gaussian with mean 0 and std deviation $\sqrt{n}$.

Law of Large Numbers:

- The average of the results obtained from a large number of trials should be close to the expected value

- **Weak Law of Large Numbers:**

$$\overline{X}_n \xrightarrow{P} \mu \qquad \text{when } n \to \infty.$$

That is, for any positive number ε,

$$\lim_{n\to\infty} \Pr\left(|\overline{X}_n - \mu| > \varepsilon\right) = 0.$$

- **Strong Law of Large Numbers:**

$$\bar{X}_n \xrightarrow{\text{a.s.}} \mu \qquad \text{when } n \to \infty.$$

That is,

$$\Pr\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1.$$

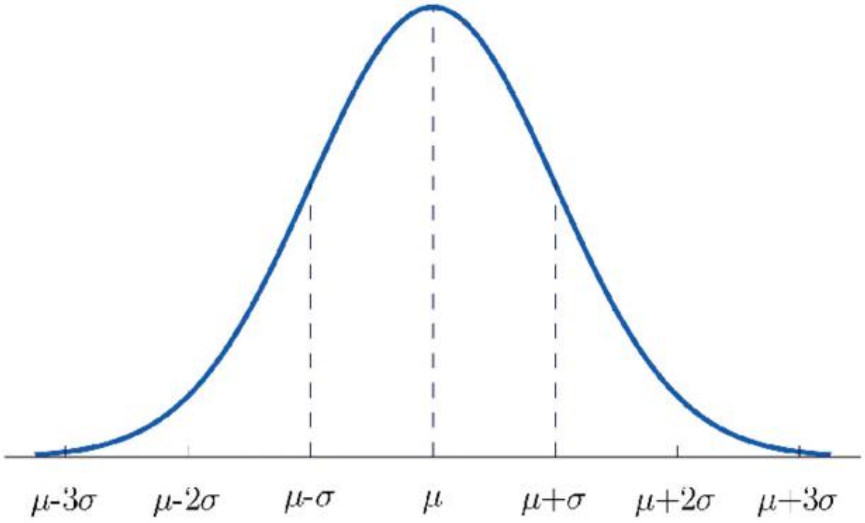
Law of Large Numbers (examples):

- A casino may lose money in a single spin of the roulette wheel; however **the house always wins** over a large number of spins.

- Roll a fair dice many times, the average of the outcomes will converge to the expectation.

- Toss a fair coin many times, the probability that the absolute difference (heads vs tails) is a small number, approaches zero as the number of flips becomes large

The normal distribution



The normal (or *Gaussian*) $N(\mu, \sigma^2)$ has mean μ, variance σ^2, and density function

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

- 68.3% of the distribution lies within one standard deviation of the mean, i.e. in the range $\mu \pm \sigma$
- 95.4% lies within $\mu \pm 2\sigma$
- 99.7% lies within $\mu \pm 3\sigma$

The central limit theorem

Suppose $X_1, \ldots, X_n$ are independent, and that they all come from the same distribution, with mean μ and variance σ^2.

Let $S_n = X_1 + \cdots + X_n$. Then S_n has mean and variance:

$$\mathbb{E}S_n = n\mu, \quad \text{var}(S_n) = n\sigma^2.$$

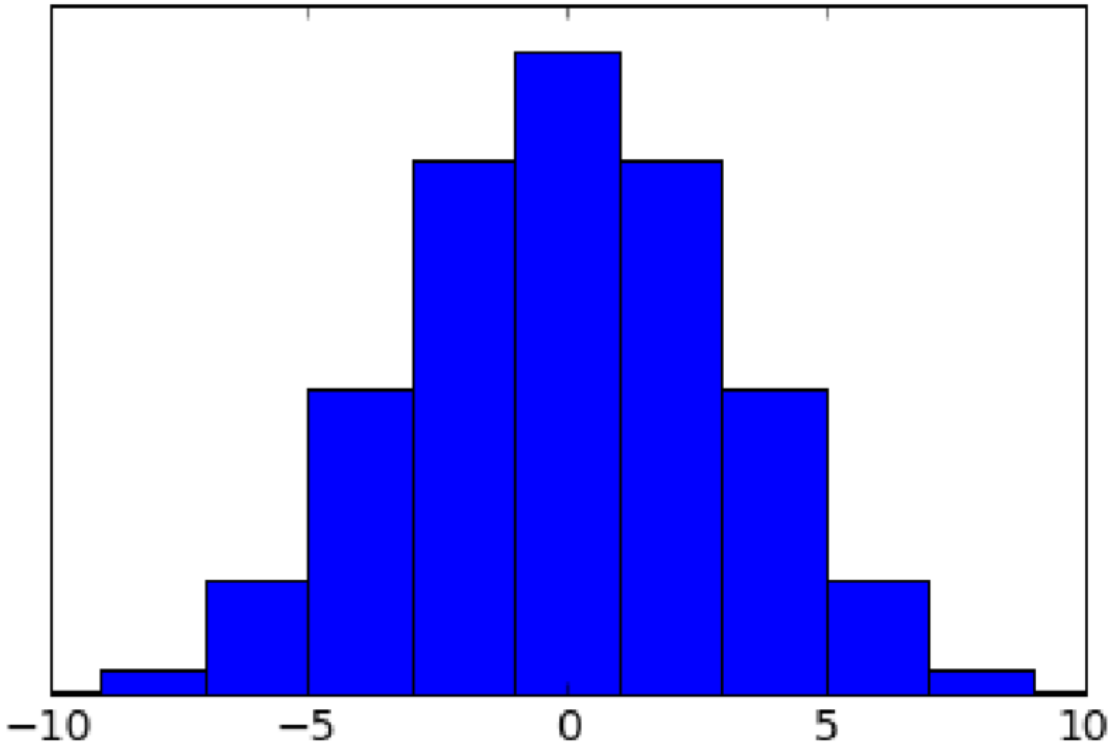
Central limit theorem, very roughly: For reasonably large n, the distribution of $S_n = X_1 + \cdots + X_n$ looks like $N(n\mu, n\sigma^2)$, the Gaussian with mean $n\mu$ and variance $n\sigma^2$.

Question: What does this imply about the average $(X_1 + \cdots + X_n)/n$? What does its distribution look like?

Answer: $N(\mu, \sigma^2/n)$.

Symmetric random walk, again

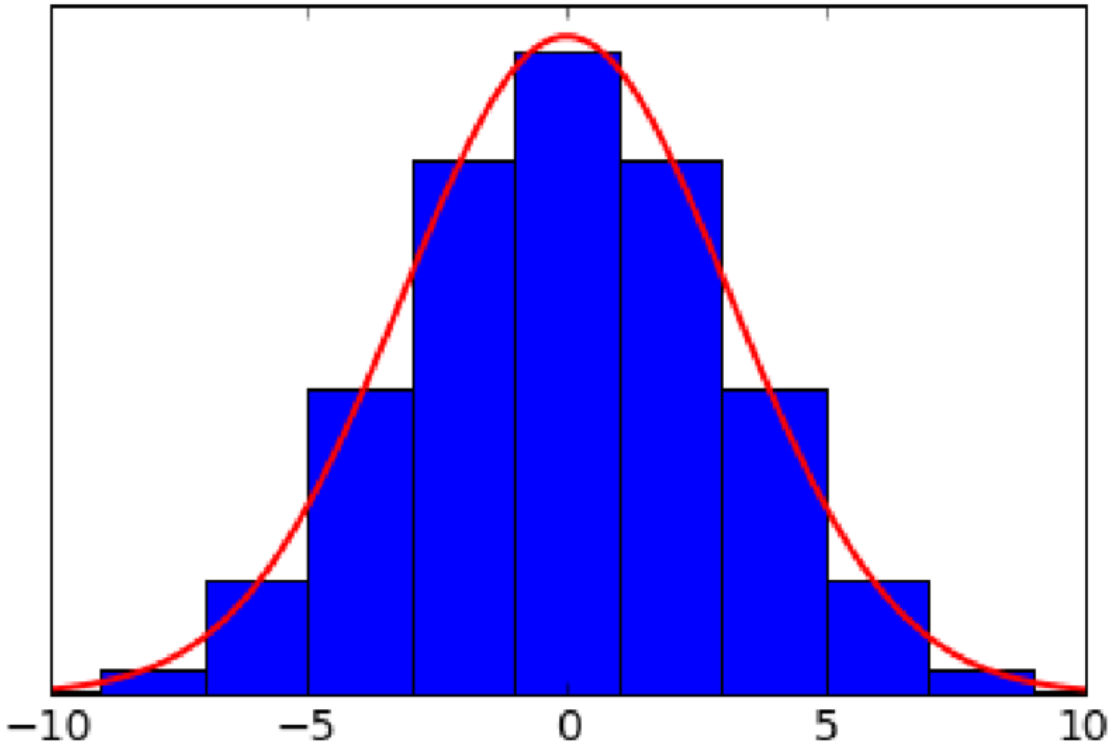
Each X_i is either 1 or -1, each with probability $1/2$.
Therefore, $X_1 + \cdots + X_n$ is distributed like $N(0, n)$.



10 steps

Symmetric random walk, again

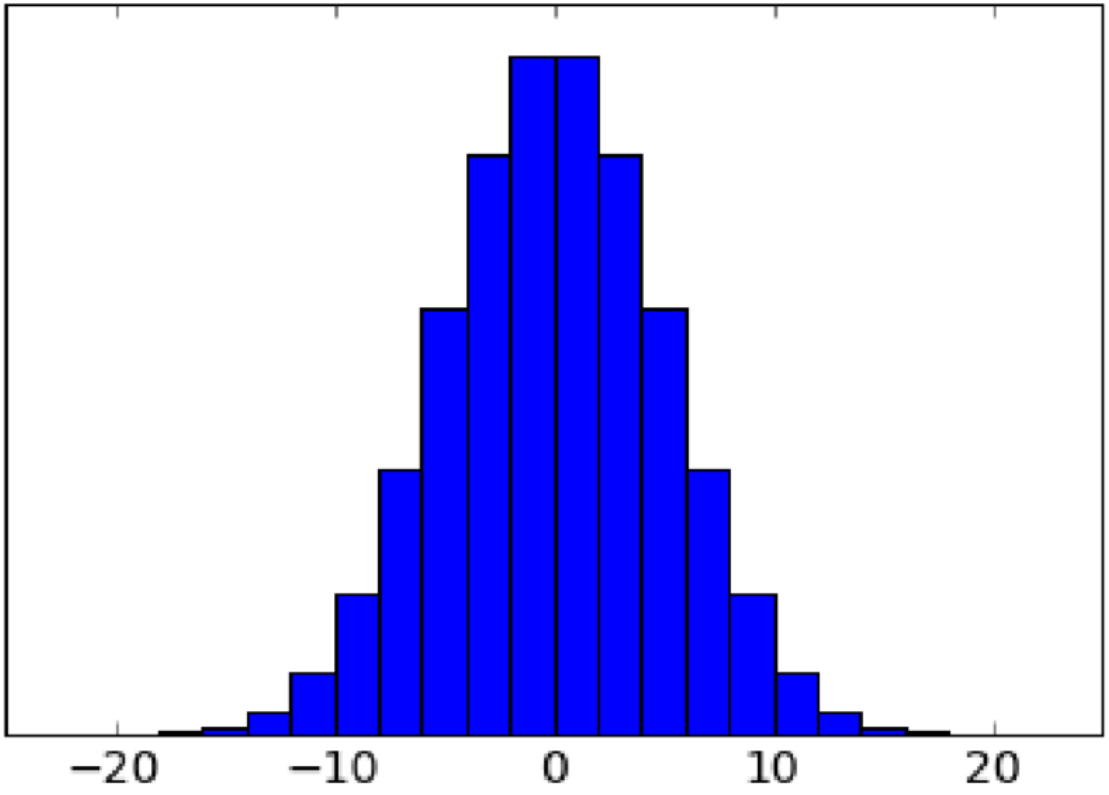
Each X_i is either 1 or -1, each with probability $1/2$.
Therefore, $X_1 + \cdots + X_n$ is distributed like $N(0, n)$.



10 steps

Symmetric random walk, again

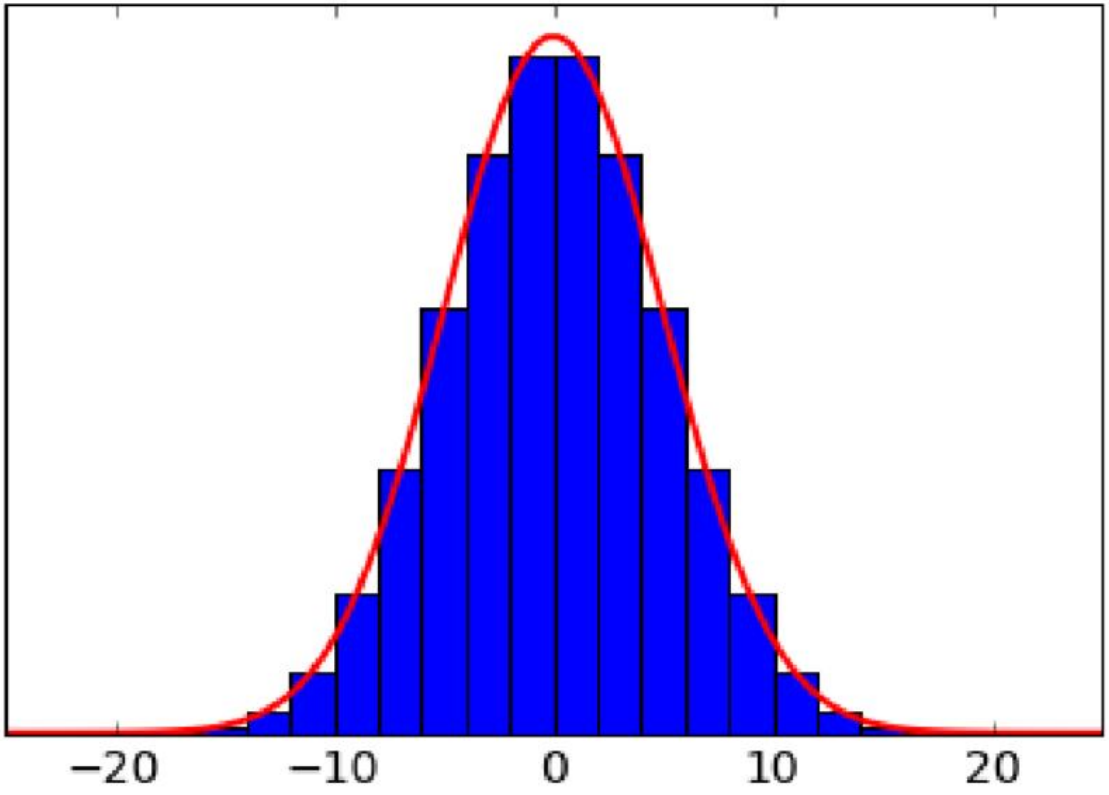
Each X_i is either 1 or -1, each with probability $1/2$.
Therefore, $X_1 + \cdots + X_n$ is distributed like $N(0, n)$.



25 steps

Symmetric random walk, again

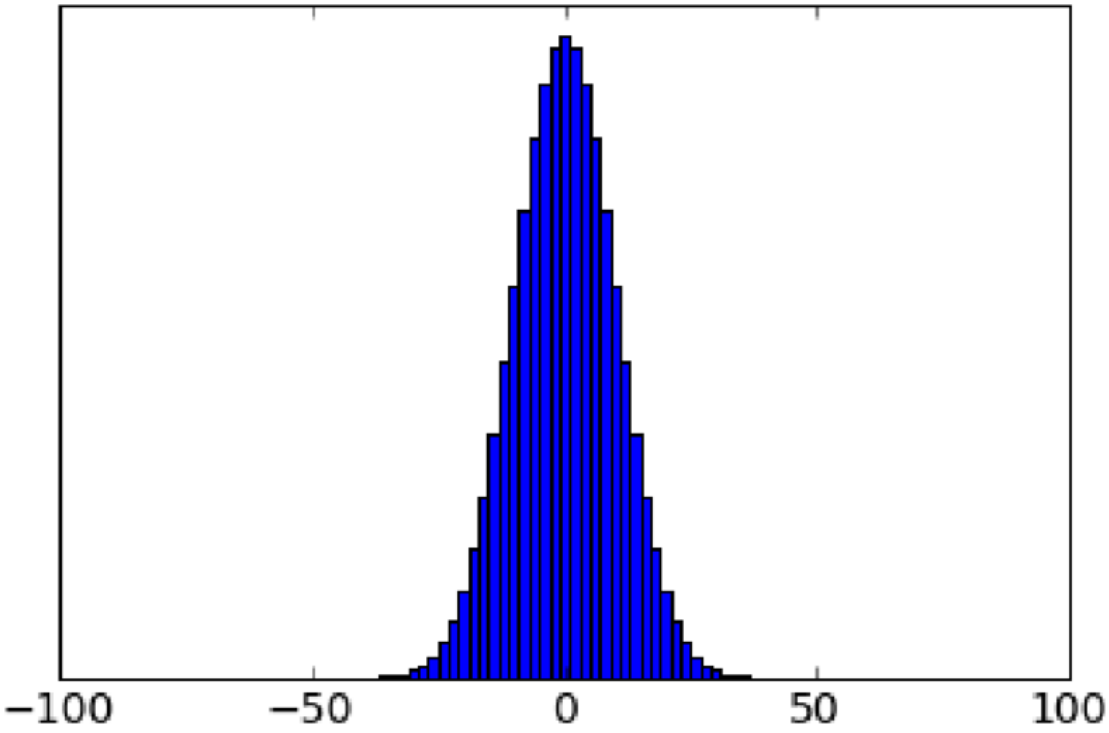
Each X_i is either 1 or -1, each with probability $1/2$.
Therefore, $X_1 + \cdots + X_n$ is distributed like $N(0, n)$.



25 steps

Symmetric random walk, again

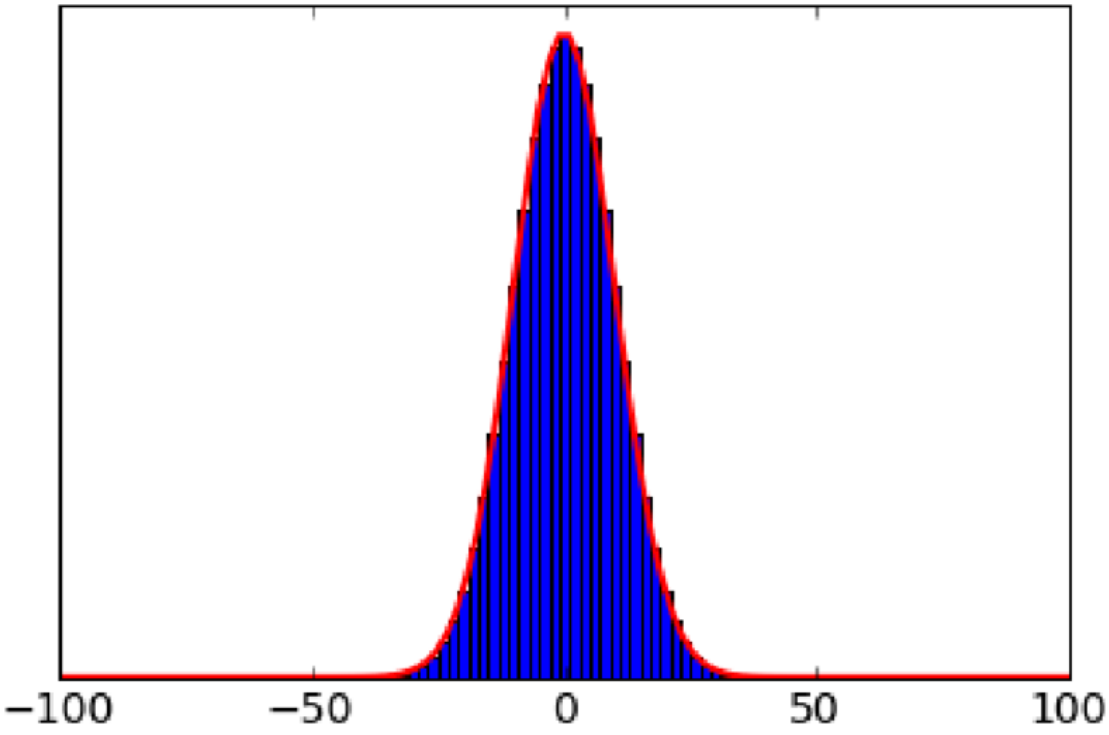
Each X_i is either 1 or -1, each with probability $1/2$.
Therefore, $X_1 + \cdots + X_n$ is distributed like $N(0, n)$.



100 steps

Symmetric random walk, again

Each X_i is either 1 or -1, each with probability $1/2$.
Therefore, $X_1 + \cdots + X_n$ is distributed like $N(0, n)$.



100 steps

Tosses of a biased coin

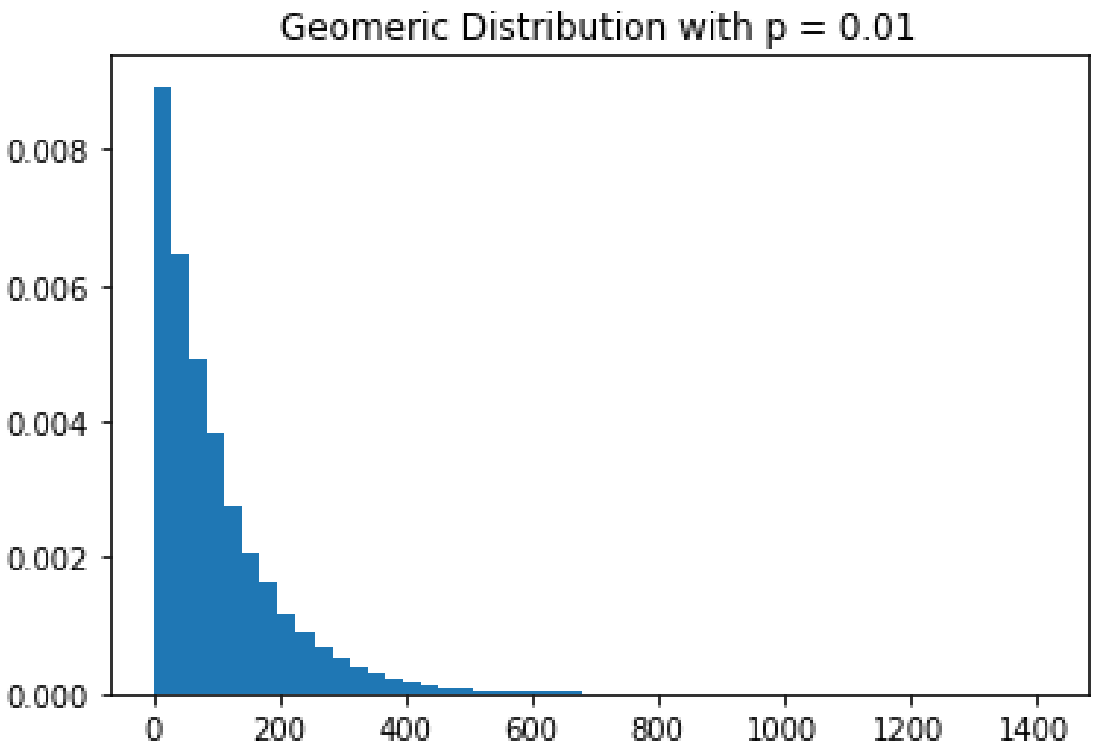
A coin of bias (heads probability) p is tossed n times.

- What is the distribution of the observed **number** of heads, roughly?
  Answer: $N(np, np(1-p))$
  Mean np, standard deviation on the order of $\sqrt{n}$.
- What is the distribution of the observed **fraction** of heads, roughly?
  Answer: $N(p, p(1-p)/n)$.
  Mean p, standard deviation on the order of $1/\sqrt{n}$.

Example: A town has 30,000 registered voters, of whom 12,000 are Democrats. A random sample of 1,000 voters is chosen. How many of them would we expect to be Democrats, roughly?

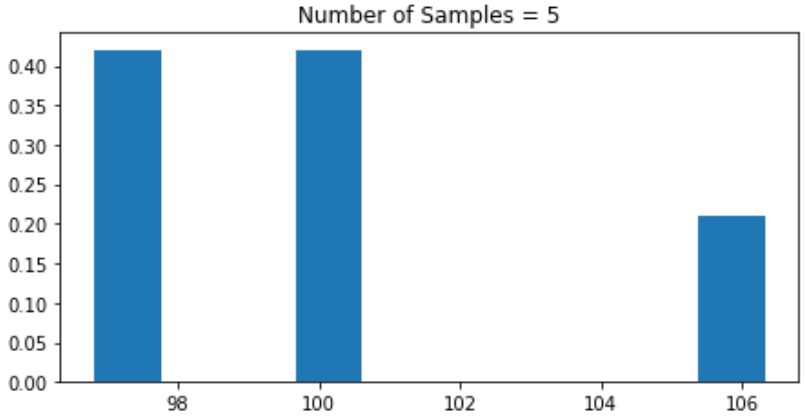
Answer: The number of Democrats observed will roughly follow a $N(1000 \times 0.4, 1000 \times 0.4 \times 0.6) = N(400, 240)$ distribution.
This has mean 400 and standard deviation ≈ 15.5.

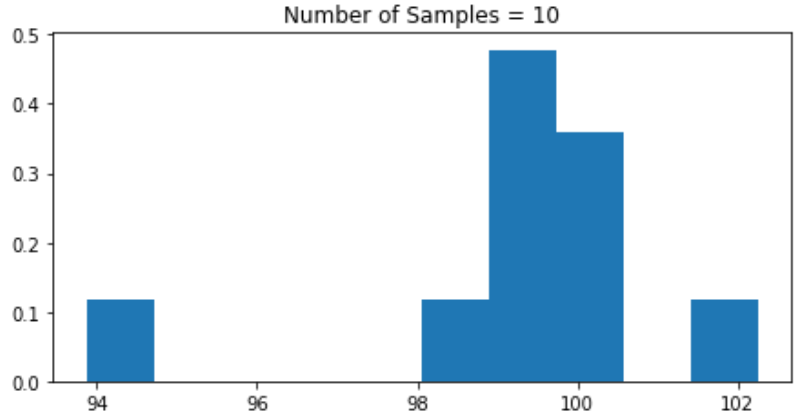
Example: Samples From Geometric Distribution vs n

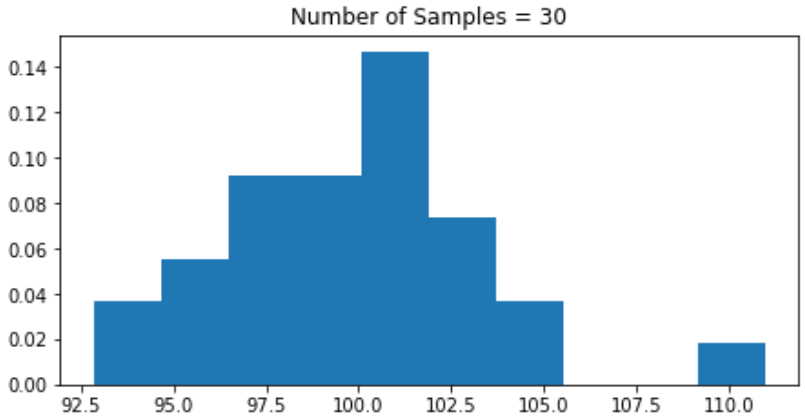


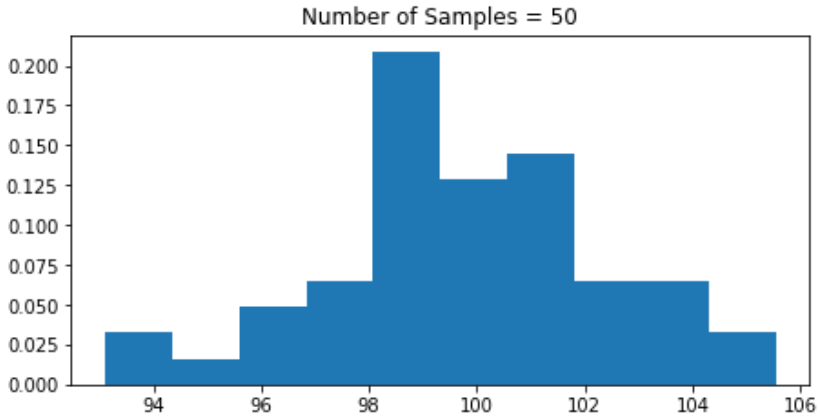
- Sample n = 5, 10, 30, 50, 100, 500 samples
- Each sample is consisted of size = 1000 instances that come from Geometric distribution
- Plot the histogram of means

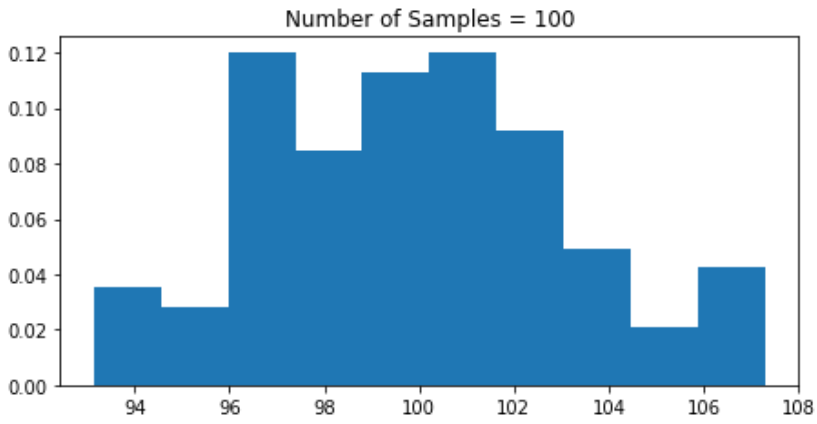
Example: Samples From Geometric Distribution vs n

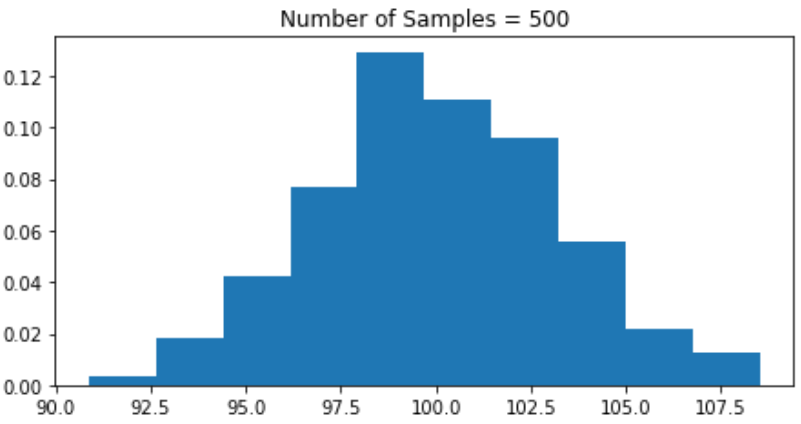




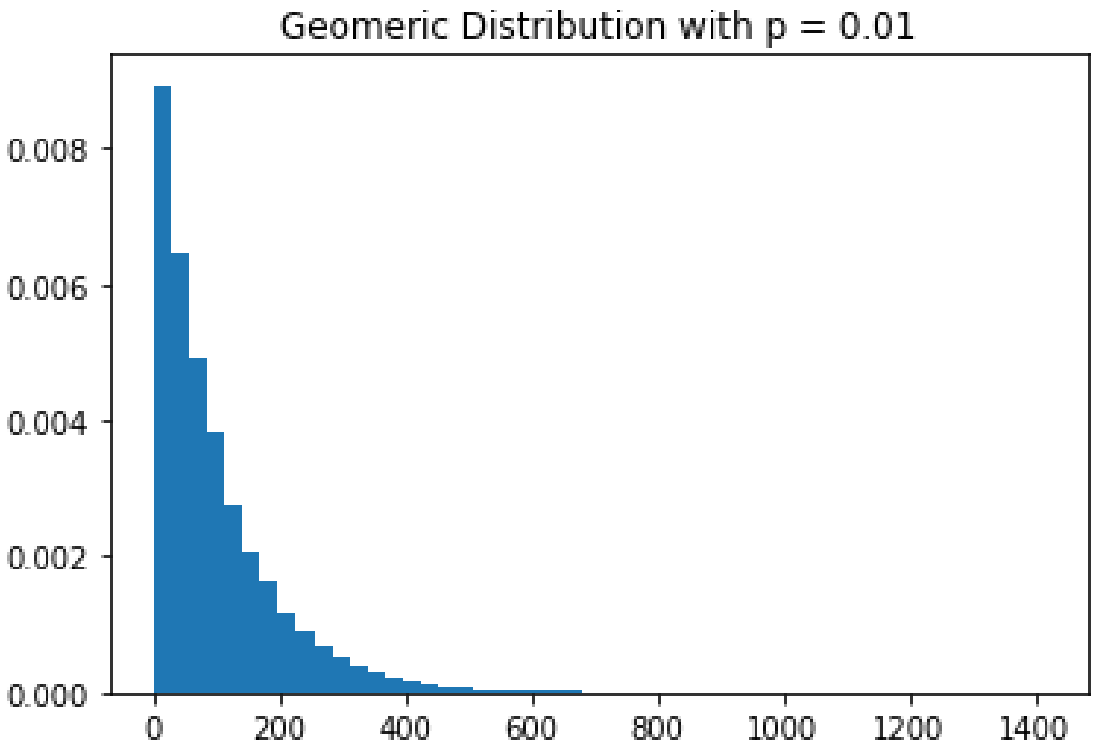






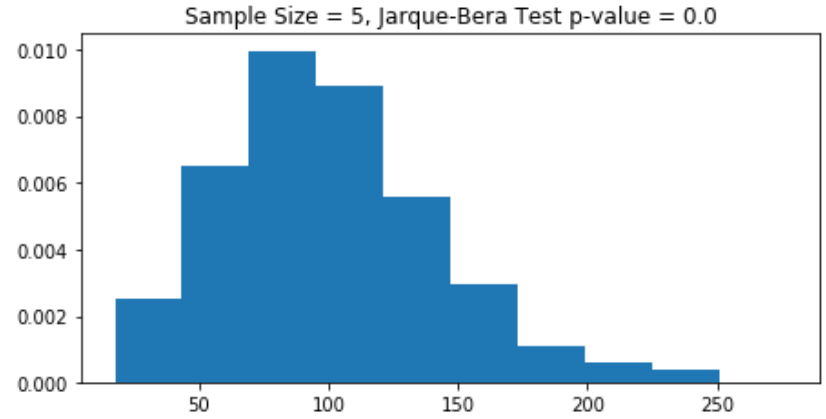


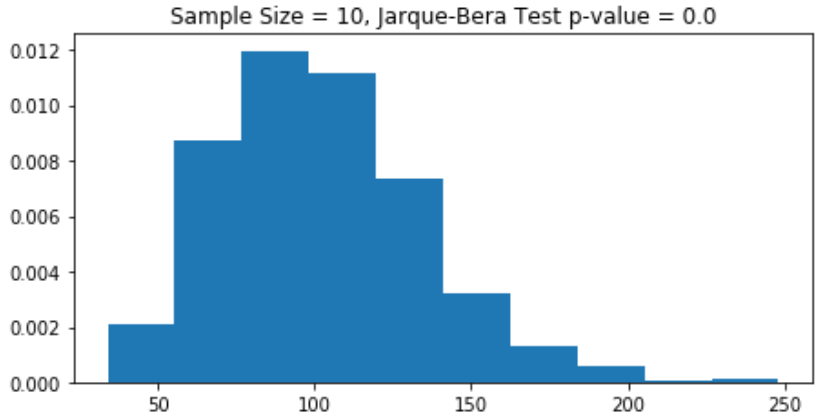
Example: Samples From Geometric Distribution vs Sample Size

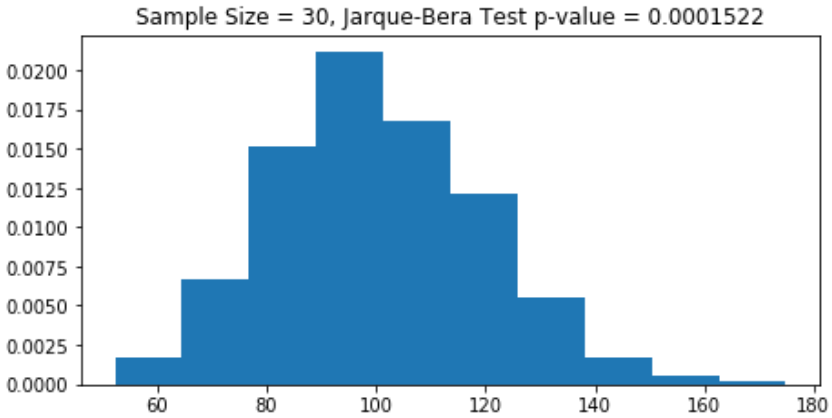


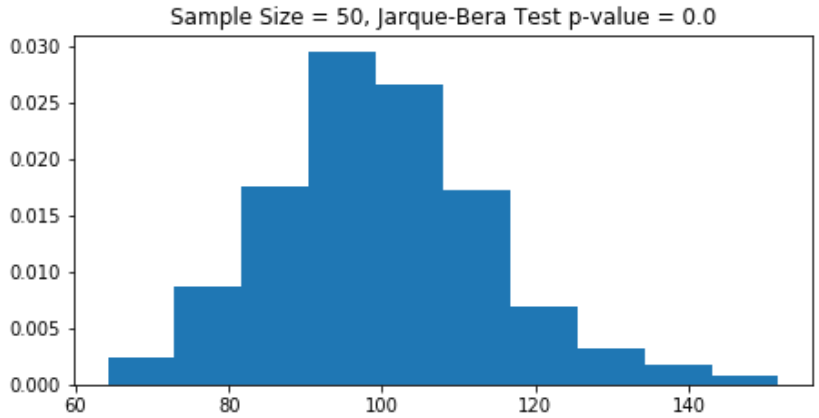
- Sample n = 1000 samples
- Each sample is consisted of size = 5, 10, 30, 50, 100, 500 instances that come from Geometric distribution
- Plot the histogram of means

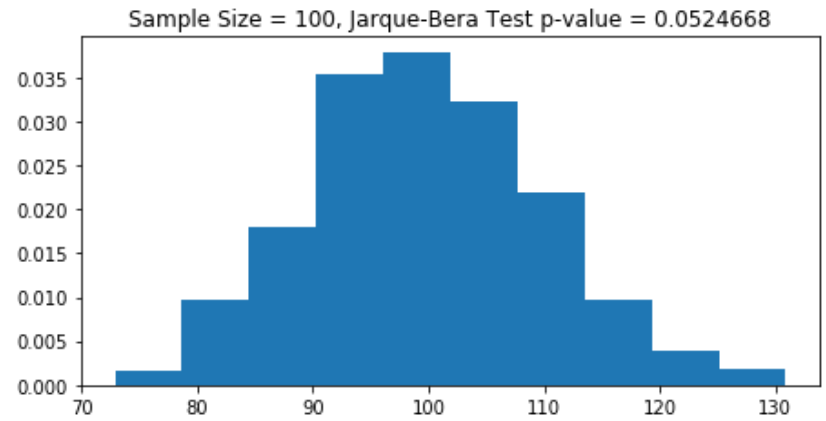
Example: Samples From Geometric Distribution vs Sample Size

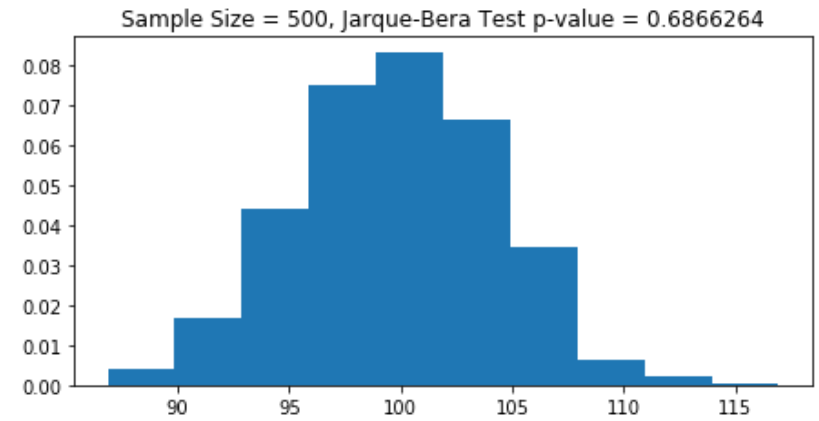




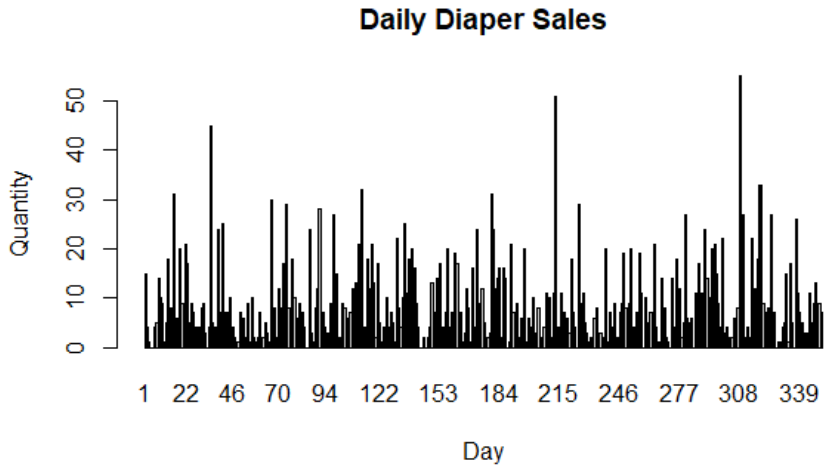


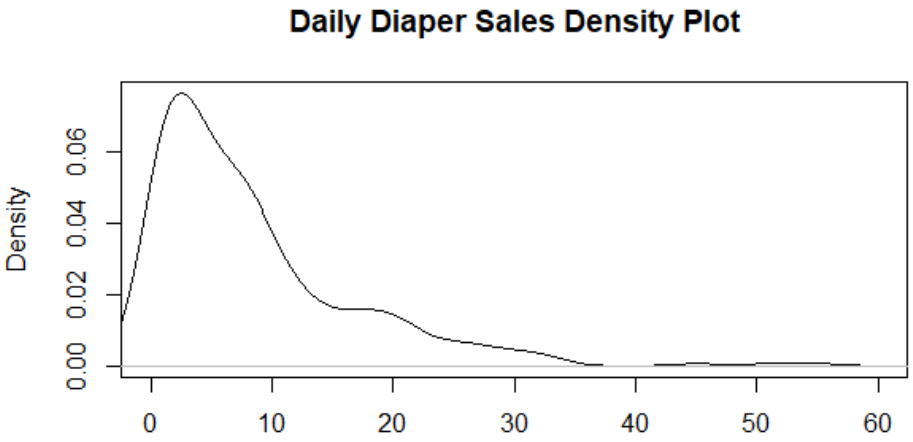


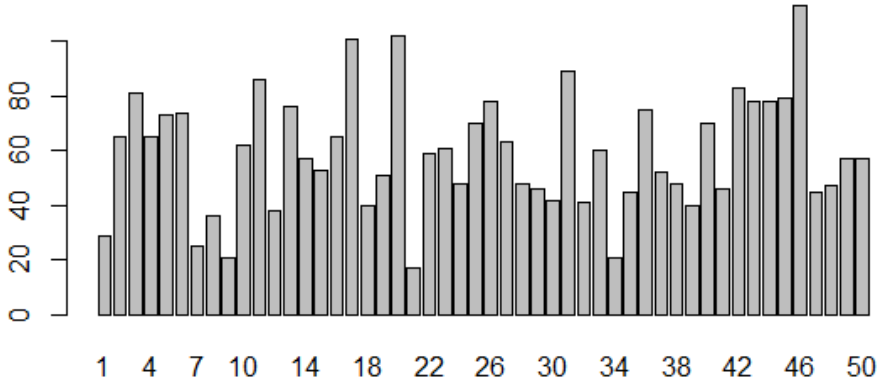


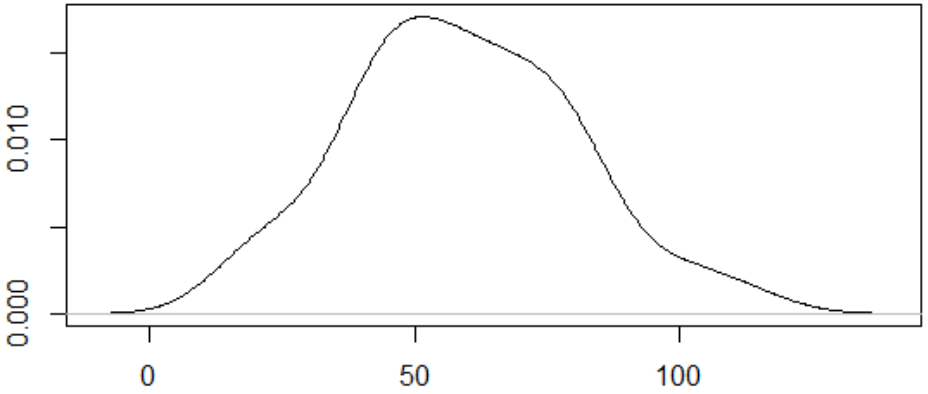


Example: Diaper Sales in a store

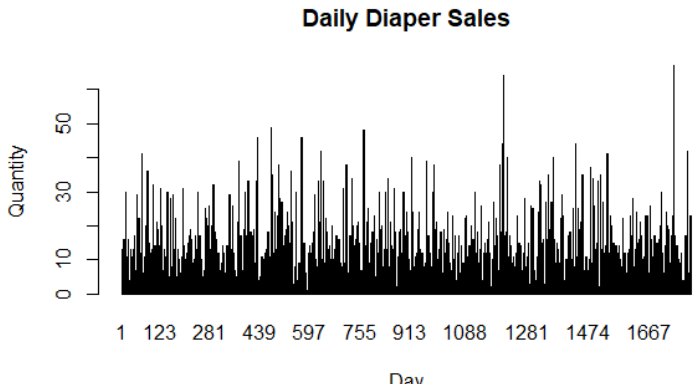


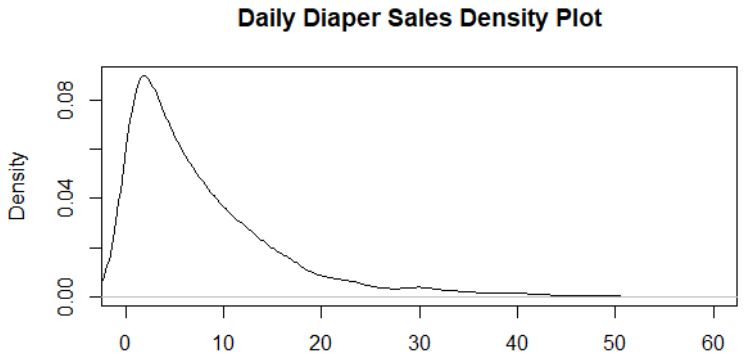


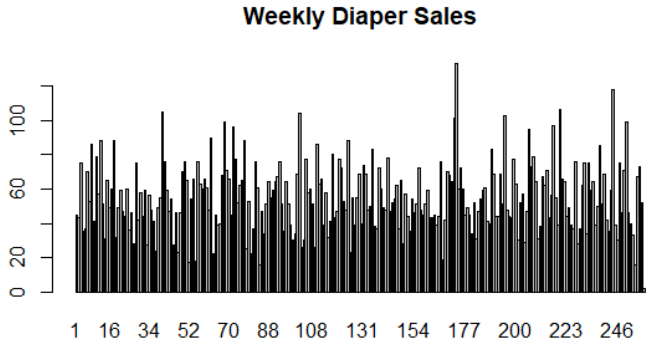


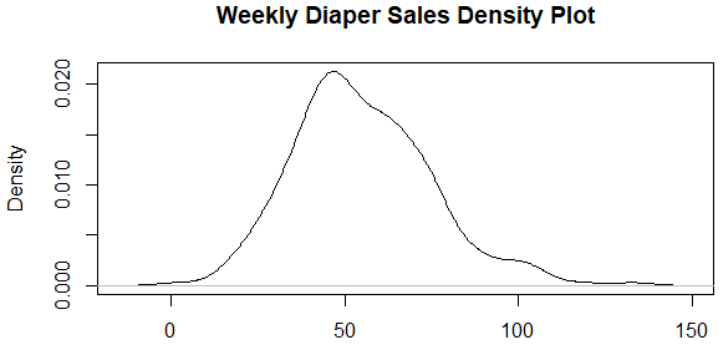


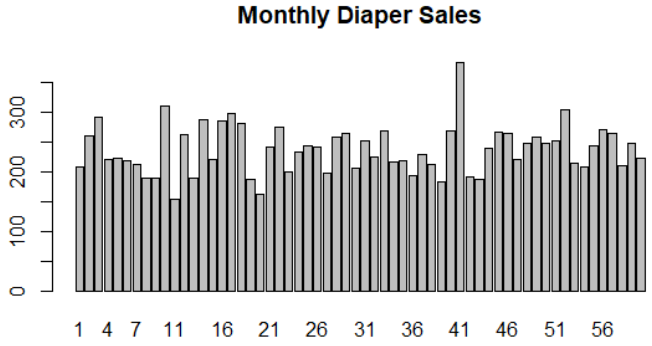
Example: Diaper Sales in a store over the years

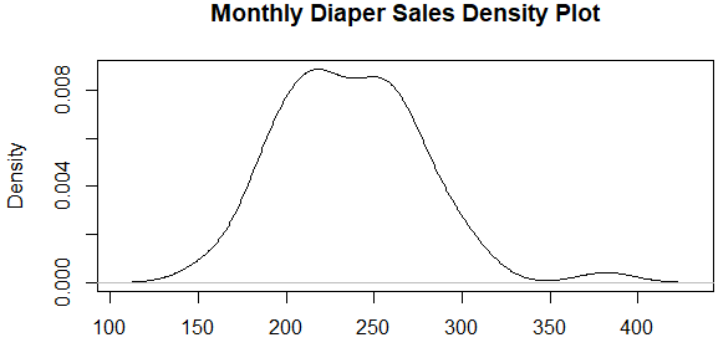












Sampling design

In the 1948 Presidential election, the polls all predicted Thomas Dewey as the winner, with at least a five-point margin. But the outcome was quite different.

Selection bias

The Republican bias in the Gallup Poll, 1936-1948.

Year	Gallup's prediction of Republican vote	Actual Republican vote
1936	44	38
1940	48	45
1944	48	46
1948	50	45

The safest way to sample is **at random**.

1936 US presidential election (The Literary Digest poll)

- In 1936 election, the candidates were the incumbent President Franklin Roosevelt (Democrat) and Alf Landon (Republican).
- *The Literary Digest* was a popular and widely read weekly magazine that ran a poll to predict the winner .
- *The Literary Digest* mailed a questionnaire to 10 million people (readers of *The Literary Digest*, registered car owners and people listed in the phone book).
- The response rate was 24%. *The Literary Digest* claimed, 'The country will know to within a fraction of one per cent the actual popular vote of forty million.'

Multistage cluster sampling

Sometimes random sampling is inconvenient, and careful multistage procedures need to be used.

For instance,

1. Stage 1
   - Divide the US into four geographical regions: Northeast, South, Midwest, West.
   - Within each region, group together all population centers of similar sizes. E.g. All towns in the northeast with 50-250 thousand people.
   - Pick a random sample of these towns.
2. Stage 2
   - Divide each town into wards, and each ward into precincts.
   - Select some wards at random from the towns chosen earlier.
   - Select some precincts at random from among these wards.
   - Then select households at random from these precincts.
   - Then select members of the selected households at random, within the designated age ranges.

Sample size versus population size

A certain town in Illinois has the same balance of Democrats and Republicans as the nation at large. We want to determine these fractions using a random sample of 1000 people. Would it be better to choose the 1000 people from the town in Illinois, or from the entire country?

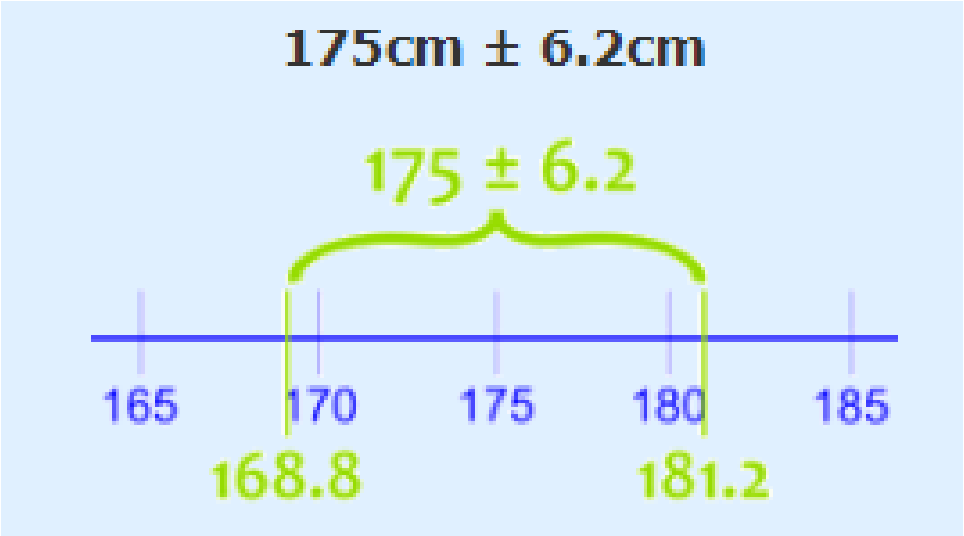
Let the unknown fraction be p. In both cases, the observed fraction will follow the $N(p, p(1-p)/1000)$ distribution.

What matters is the sample size, not the overall population size.

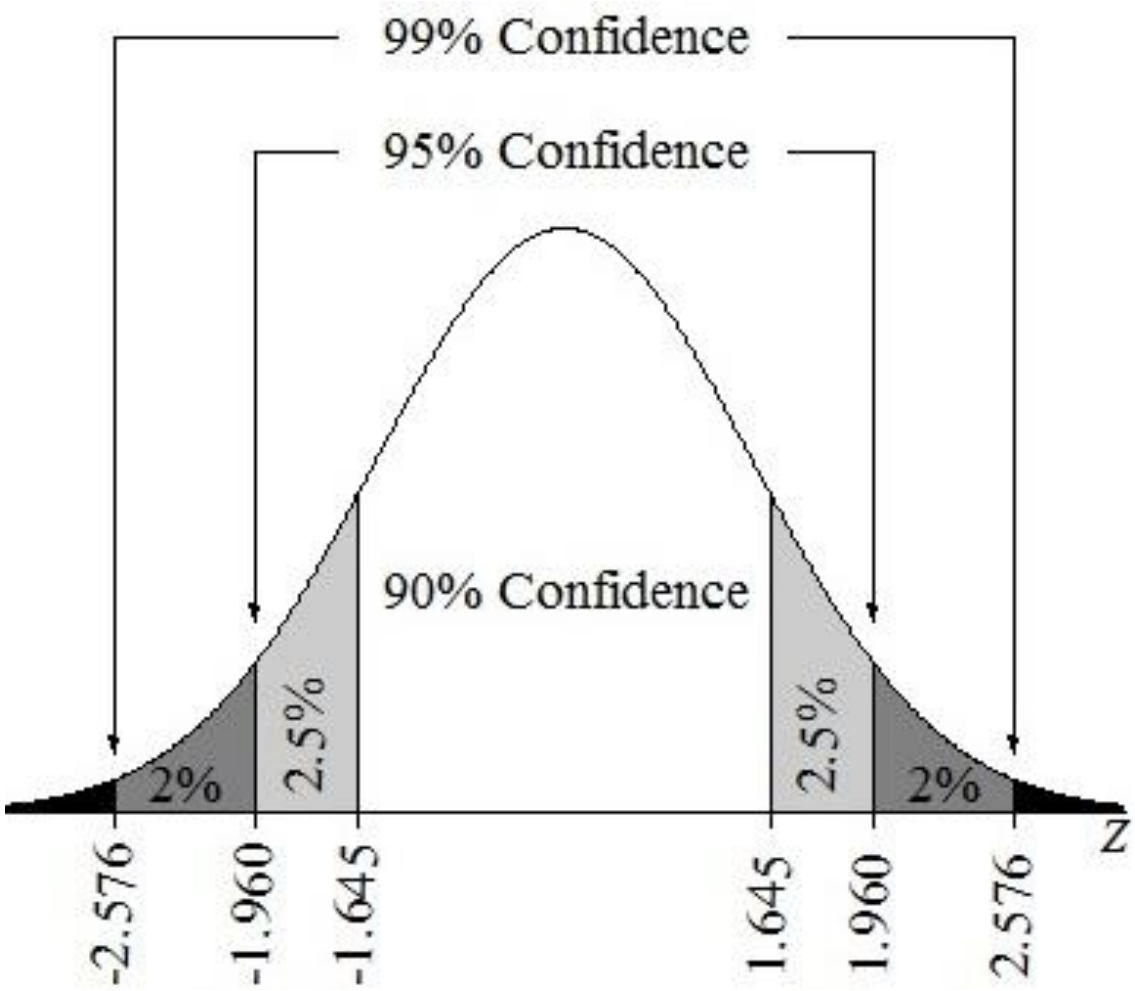
Confidence Interval:

- Statisticians use a confidence interval to describe the amount of uncertainty associated with a sample estimate of a population parameter.

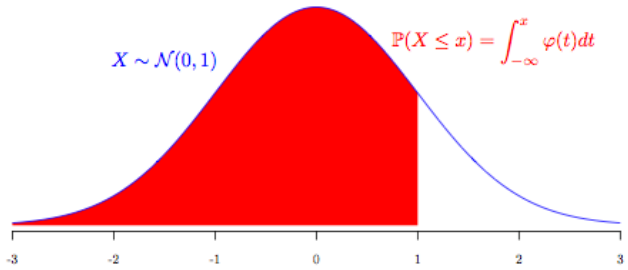
- Example – Average Male Height:
  - We measure the heights of 40 randomly chosen men, and get a mean height of 175cm,
  - We also know the standard deviation of men's heights is 20cm.
  - The **95% Confidence Interval:**



Confidence Interval:



Z Table For Normal Distribution:



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Normal Distribution:

- Standardizing
  - Find the probability from the associated X value
  - $Z = (X - \mu_X) / \sigma_X$
- Reverse Standardizing
  - Find the X value from the associated probability
  - $X = \mu_X + Z\sigma_X$

Normal Distribution:

- Example:
  - The annual snowfall in Boston is assumed to be normal with mean μ= 60 inches and standard deviation σ= 20 inches
  - What is the probability that this year's snowfall will be at least 80 inches? (Hint: $P(Z<1) = 0.8431$)
  - How much snowfall do we need this year so that this year is in top 10%? (Hint: $P(Z<1.285) \sim 0.9$)

Confidence Interval:

C	z^*
99%	2.576
98%	2.326
95%	1.96
90%	1.645

- Basic steps:

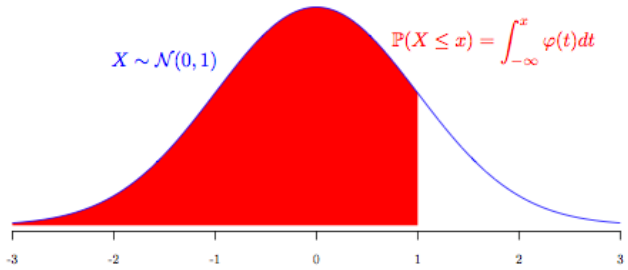
1. Identify the sample mean, $\bar{x}$
2. Identify whether the population standard deviation is known, σ, or is unknown and is estimated by the sample standard deviation s
   - If the population standard deviation is known, then:

$$z^* = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = -\Phi^{-1}\left(\frac{\alpha}{2}\right)$$

   where C = 100(1 – α)% is the confidence level and Φ is the CDF of the standard normal distribution
   - If the population standard deviation is unknown, then the Student's t distribution is used as the critical value.

3. Plug the found values into the appropriate equations:
   - For a known standard deviation: $\left(\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}}\right)$
   - For an unknown standard deviation: $\left(\bar{x} - t^* \frac{s}{\sqrt{n}}, \bar{x} + t^* \frac{s}{\sqrt{n}}\right)$

Z Table For Normal Distribution:



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

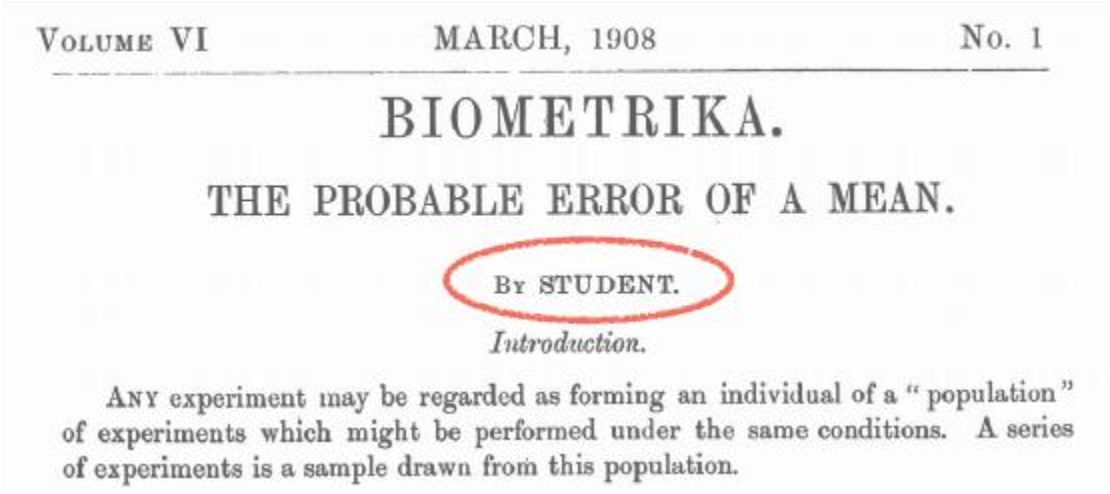
Student's t-Distribution: $T := \frac{\bar{X} - \mu}{S/\sqrt{n}}.$

- William Gossett computed the distribution of the t-statistic while working for the Guinness brewery, trying to choose the best yielding barley variety—he was concerned with small sample sizes.

- He published it under the pseudonym Student, as it was deemed confidential information by the brewery.

- The t-distribution has a single parameter called the number of degrees of freedom—this is equal to the sample size minus 1

- For large samples, typically more than 50, the sample variance is very accurate.



VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

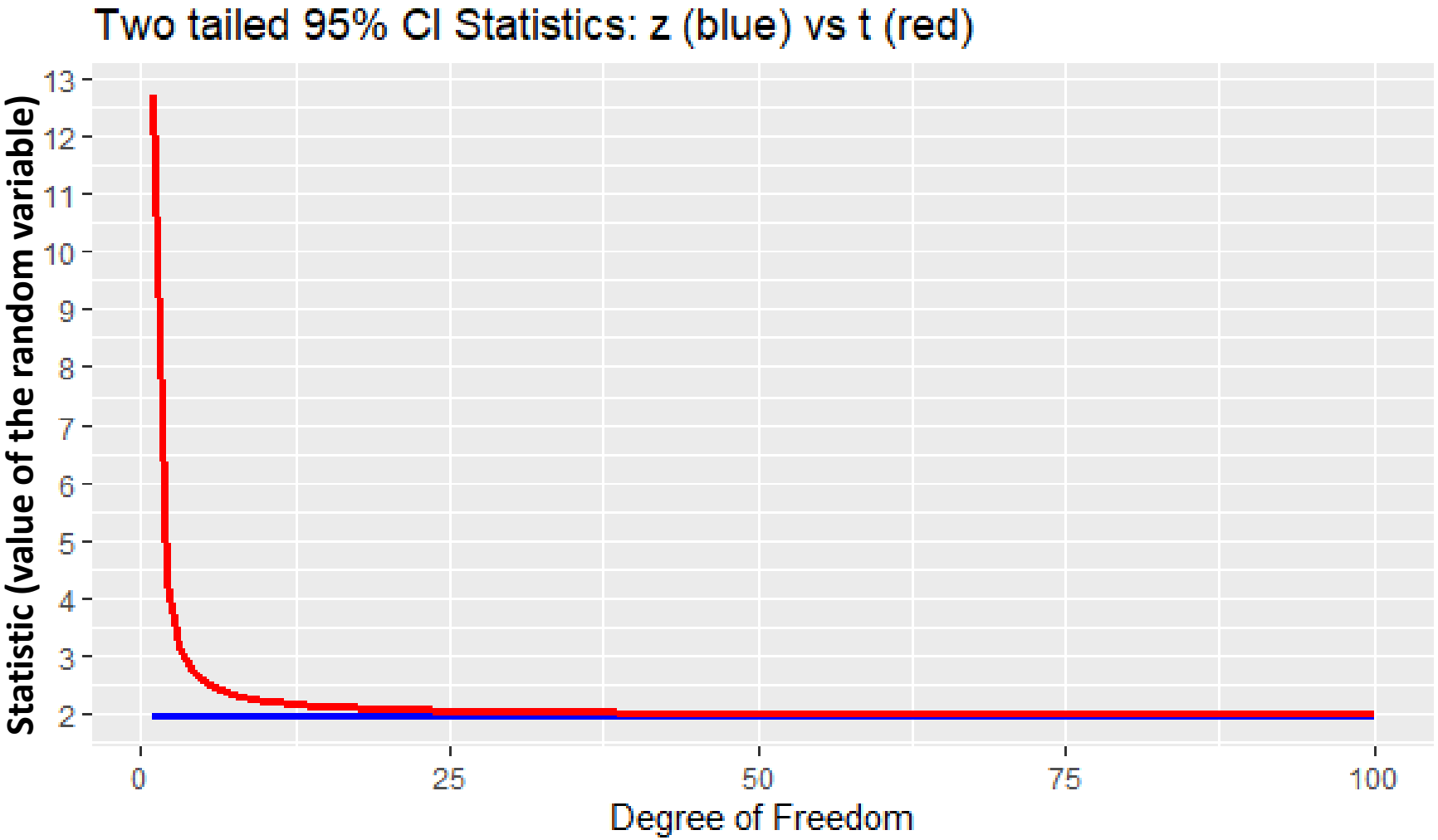
ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Degrees of freedom

- Degrees of freedom are an integral part of inferential statistical analyses, which estimate or make inferences about population parameters based on sample data. In a calculation, degrees of freedom is the number of values which are free to vary.

- As an illustration, think of people filling up a 30-seat classroom. The first 29 people have a choice of where they sit, but the 30th person to enter can only sit in the one remaining seat. Similarly, if you calculated the mean of a sample of 30 numbers, the first 29 are free to vary but 30th number would be determined as the value needed to achieve the given sample mean. Therefore, when estimating the mean of a single population, the degrees of freedom is 29.

Example: z statistic vs t statistic:



Example: estimating a fraction

A university has 25,000 registered students. In a survey, 400 students were chosen at random, and it turned out that 317 of them were living at home. Estimate the fraction of students living at home.

The observed fraction, out of $n = 400$ samples, is

$$\widehat{p} = \frac{317}{400} \approx 0.79.$$

Give error bars on this estimate.

Let p be the fraction of students living at home. Then:

$$\widehat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

Therefore, $\widehat{p}$ has standard deviation $\sqrt{p(1-p)/n}$.

But we don't know p... so what error bar to use?

In a survey, $n = 400$ students were chosen at random, and it turned out that 317 of them were living at home.

The observed fraction living at home is $\widehat{p} = 0.79$. This value $\widehat{p}$ is normally distributed with mean p and standard deviation $\sqrt{p(1-p)/n}$.

Since we don't know the true standard deviation $\sqrt{p(1-p)}$ of each sample, use the observed standard deviation $\sqrt{\widehat{p}(1-\widehat{p})}$.

$$\text{stddev}(\widehat{p}) \approx \sqrt{\frac{0.79 \times 0.21}{400}} \approx 0.02.$$

Using normal approximation gives confidence intervals:

- 68.3% interval: 0.79 ± 0.02
- 95.5% interval: 0.79 ± 0.04
- 99.7% interval: 0.79 ± 0.06

What does a 95% confidence interval mean?
It means that if we were to do this over and over again, the interval would be correct (contain the true value) at least 95% of the time.

Estimating an average

In a certain town, a random sample is taken of 400 people age 25 and over. The average years of schooling of this sample is 11.6 years, with a standard deviation of 4.1. Find a 95% confidence interval for the average educational level of people 25 and over in this town.

What is the distribution of the observed average?

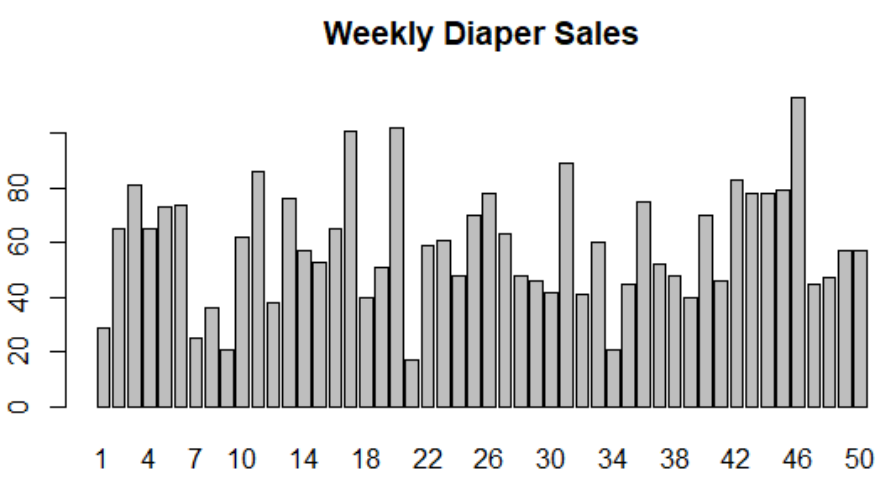
- Let the true mean educational level be μ, with stddev σ.
- We draw n samples from this distribution, and take the average $\widehat{\mu}$.
- This $\widehat{\mu}$ has distribution $N(\mu, \sigma^2/n)$.

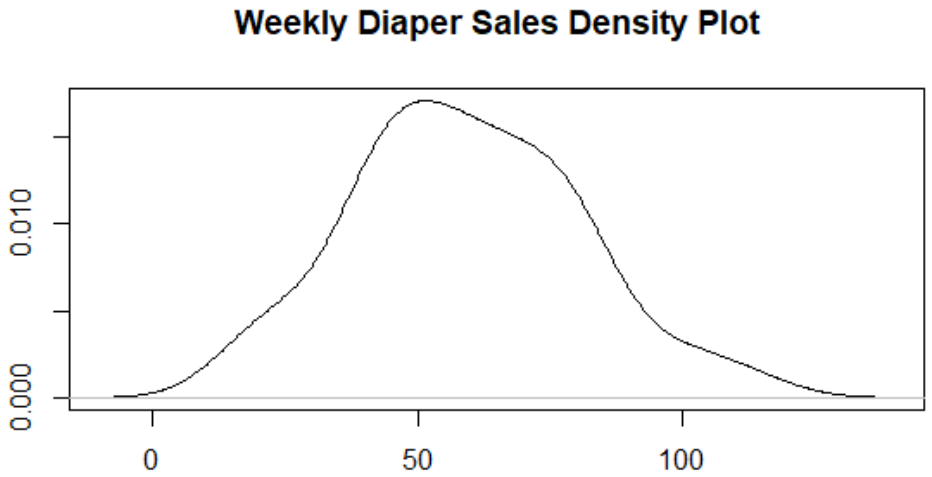
Estimate the standard deviation of $\widehat{\mu}$.

- Its standard deviation is $\sigma/\sqrt{n}$.
- We don't know σ. Instead use the sample standard deviation, 4.1.
- Standard deviation of $\widehat{\mu}$ is roughly $4.1/\sqrt{400} \approx 0.2$.

Therefore, 95% confidence interval is 11.6 ± 0.4.
And recall: the chance is in the measuring procedure, not in the quantity being estimated.

Example: Diaper Sales in a store



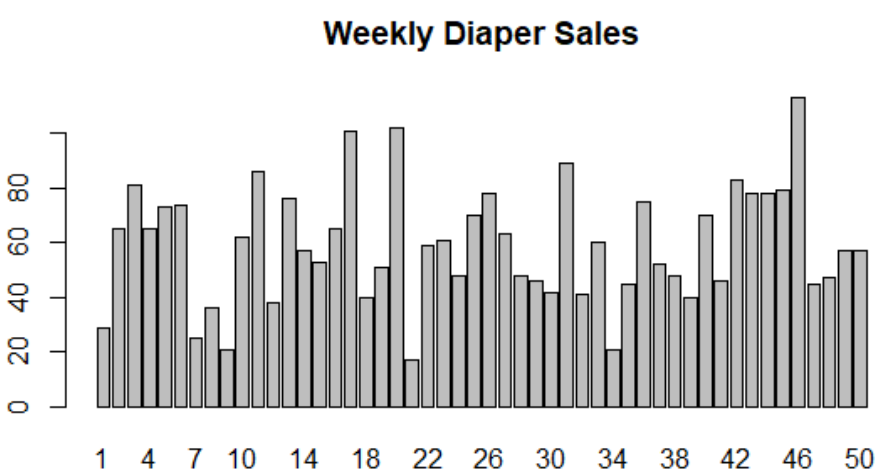


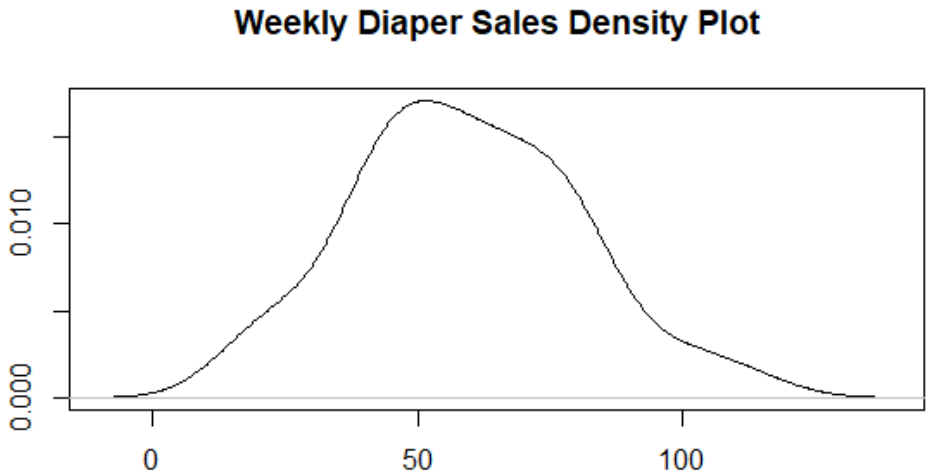
How can we construct a 95% confidence interval for the expected weekly sales?

- Using the 50 weeks of data, we found that average weekly sales quantity is 59.12, and standard deviation is 21.41.
- Assuming sample standard deviation is equal to the population standard deviation, we can construct a 95% CI for the true mean:

$$\left(\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}}\right)$$ = (53.19, 65.05) where z* is 1.96

Example: Diaper Sales in a store

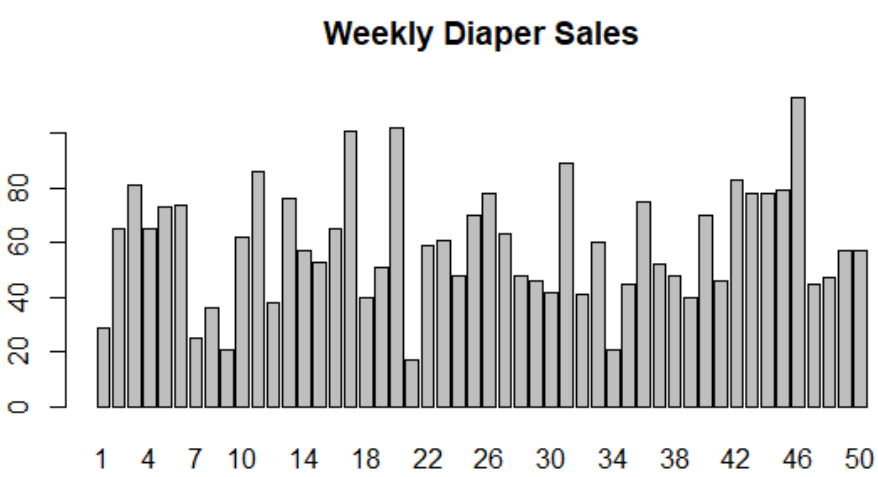


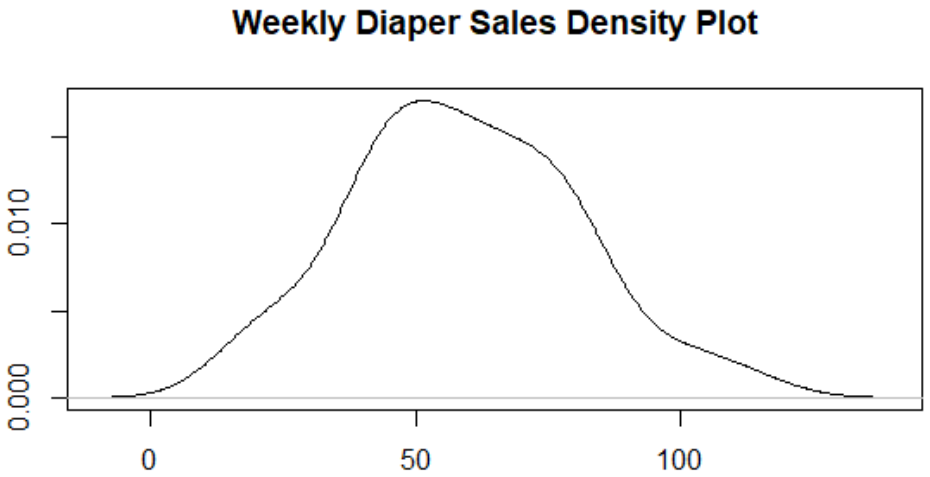


Assuming that mean and standard deviation are fixed, what will happen if we increase n?

- If n was 200:
- $\left(\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}}\right)$ = (56.15, 62.09) where z* is 1.96
- Confidence interval is narrower.

Example: Diaper Sales in a store, t-test





How can we construct a 95% confidence interval for the expected weekly sales using t-distribution?

- Using the 50 weeks of data, we found that average weekly sales quantity is 59.12, and standard deviation is 21.41.
- Using t-test we can construct a 95% CI for the true mean:

$$\left(\bar{x} - t^* \frac{s}{\sqrt{n}}, \bar{x} + t^* \frac{s}{\sqrt{n}}\right)$$ = (53.04, 65.20) where t* is 2.01 for d.o.f = 49

Kidney example: z statistic:

The level of phosphate, in mg/dl in the blood of a patient undergoing dialysis treatment was measured on six consecutive visits.

$$5.6, 5.1, 4.6, 4.8, 5.7, 6.4.$$

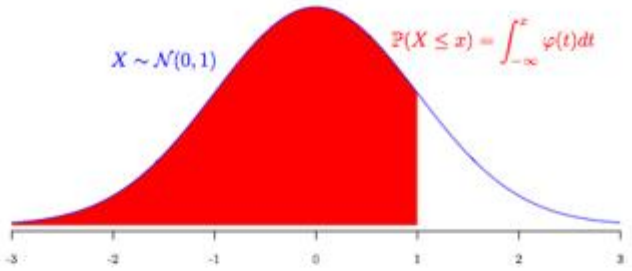
Assume that the population standard deviation is 0.6. Construct a 99% confidence interval using z-distribution.

$$\bar{x} = \frac{1}{6}(5.6 + 5.1 + 4.6 + 4.8 + 5.7 + 6.4) = 5.4\text{mg/dl}$$

The 99% Confidence interval will be:

$$\left(5.4 - z\frac{0.6}{\sqrt{6}}, 5.4 + z\frac{0.6}{\sqrt{6}}\right)$$

Kidney example: z statistic:



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.9949	0.9951	0.9952

$$\left(5.4 - z\frac{0.6}{\sqrt{6}}, 5.4 + z\frac{0.6}{\sqrt{6}}\right)$$

- z is the statistic for 99% C.I
- To construct a 99% C.I we need to cover 99% of the z-distribution. This leaves us with 0.5% probabilities on tails.
- Thus, we need to find z value which gives 99.5% cumulative probability.
- z = 2.575

Kidney example: t statistic:

The level of phosphate, in mg/dl in the blood of a patient undergoing dialysis treatment was measured on six consecutive visits.

$$5.6, 5.1, 4.6, 4.8, 5.7, 6.4.$$

Construct a symmetric 99% confidence interval.

The sample size is $n = 6$. We can compute the sample mean and sample variance as follows:

$$\bar{x} = \frac{1}{6}(5.6 + 5.1 + 4.6 + 4.8 + 5.7 + 6.4) = 5.4\text{mg/dl}$$

$$\begin{aligned} s^2 &= \frac{1}{5}(5.6 - 5.4)^2 + (5.1 - 5.4)^2 + (4.6 - 5.4)^2 \\ &\quad + (4.8 - 5.4)^2 + (5.7 - 5.4)^2 + (6.4 - 5.4)^2 \\ &= (0.67\text{mg/dl})^2. \end{aligned}$$

Kidney example: t statistic:

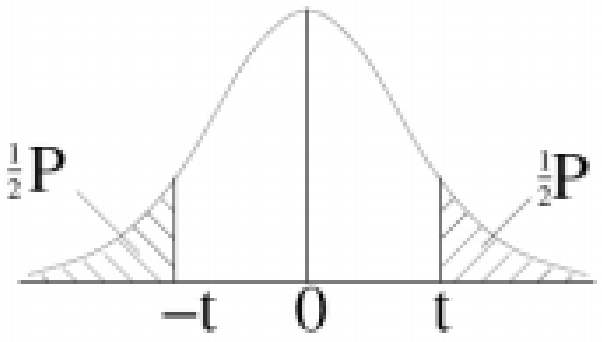
The number of degrees of freedom is $n - 1 = 5$.
Thus, the symmetric confidence interval will be

$$\left(5.4 - t\frac{0.67}{\sqrt{6}}, 5.4 + t\frac{0.67}{\sqrt{6}}\right) \text{mg/dl},$$

where t is chosen so that the T variable with 5 degrees of freedom has probability 0.01 of being bigger than t.

Using the t-Table

Two tailed t-distribution table:



Probability P of lying outside $\pm t$

d.f.	P=0.10	P=0.05	P=0.02	P=0.01
1	6.31	12.71	31.82	63.7
2	2.92	4.30	6.96	9.93
3	2.35	3.18	4.54	5.84
4	2.13	2.78	3.75	4.60
5	2.02	2.57	3.36	4.03
6	1.94	2.45	3.14	3.71

- We want to find t so that $P(-t \leq T \leq t) = 0.99$
- We look at the official table and locate the row corresponding to 5 degrees of freedom.
- The critical value is in the final column and is $t = 4.03$

Experimental design and hypothesis testing

DSE 210

A vaccine against polio

Timeline:

- 1916: First polio epidemic hit the US
- Over the next 40 years: hundreds of thousands of fatalities, especially children
- By the 1950s: several vaccines against polio were proposed
- 1954: Public Health Service and National Foundation for Infantile Paralysis (NFIP) were ready for real-world testing of a vaccine developed by Jonas Salk

How could this testing be done?

Salk vaccine: experimental design

Question: How about giving the vaccine to large numbers of children in 1954, and seeing if this led to a sharp drop in polio cases?

Bad idea: The incidence of polio varied from year to year. For instance, there were only half as many cases in 1953 than in 1952.

Controlled experiment:

- Need to deliberately leave some children unvaccinated: **controls**.
- Compare outcomes in the **treatment group** and the **control group**.

The NFIP experimental design:

- Chose two million children in selected school districts with high risk of polio, from the age groups most vulnerable (grades 1,2,3).
- Idea: would choose a million to vaccinate, and leave the rest unvaccinated, as controls.

The NFIP experimental design

How to partition the subjects into treatment and control groups?

- NFIP split it by grade level: grade 2 would get the vaccine, grades 1 and 3 would be controls.
- This is problematic. What if the incidence were higher in one grade than another? Such factors would **confound** the effect of treatment. Better idea: divide randomly.

A significant complication: parental consent.

- Those chosen for vaccination needed parental consent. Half the parents refused.
- Higher-income parents more likely to consent to treatment. Does this bias the study for or against the vaccine?
  Against. Children in less hygenic surroundings tend to contract mild cases while still protected by mother's antibodies, and this protects them later.

A better design

Textbook design: **randomized controlled double-blind** experiment.

- Control group needs to be from the same population as the treatment group.
  Therefore, select both from children whose parents consented to treatment.
- Choose the two groups at random from the same population.
  This is a **randomized controlled** experiment.
- Subjects should not know which group they are in.
  Therefore, children in the control group should be given a placebo.

Both designs were used: some school districts used the NFIP design, others used the double-blind design.

Salk vaccine: the results

For the double-blind randomized controlled experiment:

	Size	Rate (per 100K)
Treatment	200,000	28
Control	200,000	71
No consent	350,000	46

(The NFIP experiment showed a significantly weaker effect.)

How can we assess the significance of these numbers?

Historical controls

Sometimes, experiments compare outcomes for people receiving a new treatment to outcomes observed in the past without that treatment (historical controls). This is inferior to a randomized controlled design.

Example: What is the value of coronary bypass surgery for patients with coronary artery disease? Two studies, one with randomized controls and one with historical controls, reported these three-year survival rates:

	Randomized	Historical
Surgery	87.6%	90.0%
Controls	83.2%	71.1%

How might this discrepancy be explained?

Historical control trials can cause selection bias. Determining the groups in historical control trials is a challenging task.

Experimental Design:

Design of experiments

- Controlled experiments
- Observational studies

Observational studies

Two kinds of study:

- **Controlled experiment**: investigators decide who is in the treatment group and who is in the control group.
- **Observational study**: the subjects assign themselves to these two groups. The investigators just watch.

Example: studies on smoking are necessarily observational.

- Heart attacks, lung cancer, and various other diseases are more common among smokers than non-smokers.
- But perhaps there are other explanations: confounding factors that make people smoke and also make them sick.
- For instance: sex. Men are more likely to smoke than women, and are more likely to get heart disease.
- Or age: older people have different smoking habits and are more at risk for these diseases.

Careful observational studies have controlled for many confounding factors and together make a case that smoking does cause these diseases.

Cervical cancer and circumcision

For many years, cervical cancer was one of the most common cancers among women.

- Investigators looking for causes found that cervical cancer seemed to be rare among Jews.
- They also found it to be quite rare among Muslims.
- In the 1950s, various investigators concluded that circumcision of males protected against this cancer.

More recent studies suggest that cervical cancer is caused by human papilloma virus, which is sexually transmitted. More sexually active women, with more partners, are more likely to be exposed to it.

Ultrasound and low birthweight

Experiments on lab animals showed that ultrasound can cause low birthweight. Is this true for humans?

- Investigators at Johns Hopkins ran an observational study.
- They tried to adjust for various confounding factors.
- Even controlling for these, babies exposed to ultrasound on average had lower birthweight than those not exposed.

At that time, ultrasounds were used mostly during problem pregnancies: the common cause of the ultrasound and low birthweight. A later randomized controlled experiment showed no harm.

Statistical hypothesis testing

1. The z statistic
   - Testing the mean of a distribution
   - Testing whether two distributions have the same mean
2. The χ^2 statistic
   - Testing whether a sequence of $\{1, 2, \ldots, k\}$ outcomes comes from a particular k-sided die
   - Testing the independence of two variables

Hypothesis Testing

- A statistical hypothesis is an assertion or conjecture concerning one or more populations.

- To prove that a hypothesis is true, or false, with absolute certainty, we would need absolute knowledge. That is, we would have to examine the entire population.

- Instead, hypothesis testing concerns on how to use a random sample to judge if it is evidence that supports or not the hypothesis.

- Hypothesis testing is formulated in terms of two hypotheses:
  - H_0: the null hypothesis;
  - H_1: the alternative hypothesis.

Hypothesis Testing

The hypothesis we want to test is if H_1 is "likely" true.

So, there are two possible outcomes:

- Reject H_0 and accept H_1 because of sufficient evidence in the sample in favor of H_1;
- Do not reject H_0 because of insufficient evidence to support H_1.

Very important!!

Note that failure to reject H_0 does not mean the null hypothesis is true. There is no formal outcome that says "accept H_0." It only means that we do not have sufficient evidence to support H_1.

Example: new tax code

A senator introduces a change to the tax code that he claims is revenue-neutral. How can this be verified?

- See how this change would affect last year's tax returns.
- Pick 100 returns at random, look at the change in revenue of each.
- The average change is $-219.
- The standard deviation is $725.

Analyze this in the framework of **hypothesis testing**.

- **Null hypothesis**: The average change is $0.
- **Alternative hypothesis:** The average change is negative.

In order to discredit the null hypothesis, *argue by contradiction.*

- Assume the null is true.
- Compute a **statistic** that measures the difference between what is observed and what would be expected under the null.
- What is the chance of obtaining a statistic this extreme?

The z statistic

Pick 100 tax returns at random.

- The average change in revenue is $X = -219$ dollars.
- The standard deviation is \$725.

How likely is X under the null?

- Recall null hypothesis: expected change is \$0.
- Under the null, X would be normally distributed with mean 0 and standard deviation $725/10 = 72.5$.
- The observed X is ≈ 3 standard deviations from the mean: unlikely.

The z-**statistic** measures how many standard deviations away the observed value is from its expectation.

$$z = \frac{\text{observed} - \text{expected}}{\text{standard deviation}} = \frac{-219 - 0}{72.5} \approx -3$$

The probability of observing this under the null is the p-**value**.

This p-value is less than $1/1000$: strong evidence against the null.

Hypothesis testing: recap

The null hypothesis is what we are trying to discredit.

We do this by contradiction:

- Let the observation be denoted X.
- What is the distribution of X under the null?

If we would expect X to be normally distributed, we can use the z-statistic:

$$z = \frac{\text{observed } X - \text{expected } X}{\text{standard deviation of } X}$$

The p-value is the probability of seeing a value (at least) this extreme under the null. A small p-value is evidence against the null.

Example: an ESP demonstration

Charles Tart's experiments at UC Davis using the "Aquarius":

- Aquarius has an electronic random number generator
- Chooses one of four targets but doesn't reveal which
- The subject guesses which, and a bell rings if correct

The specific experiment:

- 15 subjects who considered themselves clairvoyant
- Each made 500 guesses, total of 7500
- Of these, 2006 were correct
- Compare to $7500/4 = 1875$

How significant is this?

ESP: analysis

Total of 7500 trials.

- Each time: one of four outcomes
- Total number of correct guesses: 2006

Null hypothesis: The data comes from a coin of bias 0.25.

Assume the null is true.

- The total number of successes in 7500 trials is approximately normal with what mean and standard deviation?

$$\text{Mean} = 7500 \times 0.25 = 1875$$

$$\text{Stddev} = \sqrt{7500 \times 0.25 \times 0.75} \approx 37$$

- The z statistic:

$$z = \frac{\text{observed} - \text{expected}}{\text{standard deviation}} \approx \frac{2006 - 1875}{37} \approx 3.5$$

This is strong evidence against the null.

Example: improving math scores?

National Assessment of Educational Progress data on 17-year olds:

- Average math score in 1978 was 300.4, with standard deviation 30.1
- Average math score in 1992 was 306.7, with standard deviation 34.9
- Both based on random sample of 1000 students

How significant was the improvement?

Null hypothesis: The means of the two distributions (scores in 1978, scores in 1992) are the same.

Assume the null is true. Let μ be the common mean.

- The sample average in 1978, call it X_1, is roughly normal with mean μ and standard deviation $\sigma_1 = 30.1/\sqrt{1000} \approx 1.0$.
- The sample average in 1992, call it X_2, is roughly normal with mean μ and standard deviation $\sigma_2 = 34.9/\sqrt{1000} \approx 1.1$.
- The difference $X_2 - X_1$ is therefore normally distributed with mean zero and standard deviation $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2} \approx 1.5$.

What is the z-statistic? And what can we conclude?

Math scores, cont'd

100 students chosen at random in 1978 and 1992. Math scores recorded.

$X_1 =$ sample average score in 1978

$X_2 =$ sample average score in 1992

Null hypothesis: The means of the two distributions (scores in 1978, scores in 1992) are the same.

Under the null, $X_2 - X_1$ is normally distributed with mean zero and standard deviation $\sigma = 1.5$.

Observed scores: $X_1 = 300.4$ and $X_2 = 306.7$.
The z-statistic for $X_2 - X_1$ is

$$z = \frac{\text{(observed)} - \text{(expected)}}{\text{standard deviation}} = \frac{306.7 - 300.4}{1.5} \approx 2.1.$$

The observed difference has probability about 2% under the null: strong evidence against the null.

Example: the influence of wording

Study by Amos Tversky. 167 doctors were given information about the effectiveness of *surgery* versus *radiation therapy* for lung cancer. The same information was presented two ways.

80 of the doctors got Form A:

> *Of 100 people having surgery, 10 will die during treatment, 32 will have died by one year, and 66 will have died by five years. Of 100 people having radiation therapy, none will die during treatment, 23 will die by one year, and 78 will die by five years.*

The other 87 doctors got Form B:

> *Of 100 people having surgery, 90 will survive the treatment, 68 will survive one year or longer, and 34 will survive five years or longer. Of 100 people having radiation therapy, all will survive the treatment, 77 will survive one year or longer, and 22 will survive five years or longer.*

At the end, each doctor was asked which therapy he or she would recommend for a lung cancer patient.

	Form A	Form B
Favored surgery	40	73
Favored radiation	40	14
Total	80	87
Fraction favoring surgery	0.50	0.84

Let p_A be the probability that a doctor reading form A favors surgery, and let p_B be the probability that a doctor reading form B favors surgery.
Null hypothesis: $p_A = p_B$.

Let X_A, X_B be the observed fractions favoring surgery.

- X_A is (roughly) normally distributed, with mean p_A and standard deviation $\sigma_A = \sqrt{(0.5 \times 0.5)/80} \approx 0.056$.
- X_B is (roughly) normally distributed, with mean p_B and standard deviation $\sigma_B = \sqrt{(0.84 \times 0.16)/87} \approx 0.039$.
- Under the null, $X_A - X_B$ is normally distributed with mean zero and standard deviation $\sigma = \sqrt{\sigma_A^2 + \sigma_B^2} \approx 0.068$.

Then $z \approx 5.0$. Very unlikely under the null!

Back to the Salk vaccine

	Size	Number of cases
Treatment	200,000	57
Control	200,000	142
No consent	350,000	92

Null hypothesis: Both groups have the same chance of getting polio.

Let X_t be the number of observed cases in the treatment group and X_c the number of observed cases in the control group.

- X_t is (roughly) normally distributed, with standard deviation $\approx \sqrt{57}$
- X_c is (roughly) normally distributed, with standard deviation $\approx \sqrt{142}$
- Under the null, $X_c - X_t$ is normally distributed with mean zero and standard deviation $\sqrt{57 + 142} \approx 14$.

The z statistic for $X_c - X_t$ is then

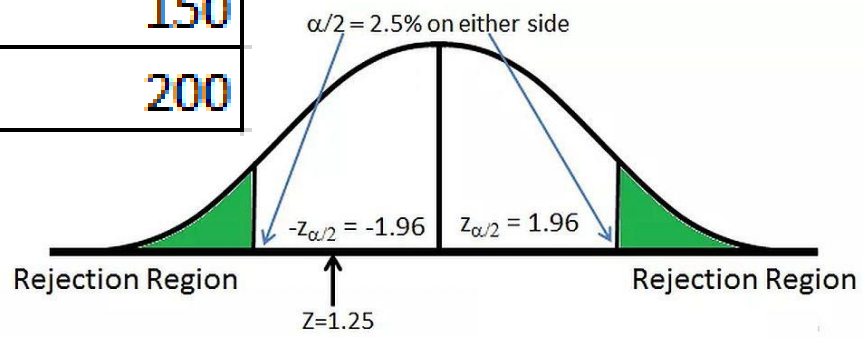
$$z \approx \frac{142 - 57}{14} \approx 6.1.$$

The observed difference is extremely unlikely under the null.

Hypothesis Testing: A/B Testing Example

An e-commerce company is running an experiment to decide on their website design. According to this test, 50% of the visitors will see the old design and the other 50% will see the new design. Can we assert that new website design changed the sales with 95% confidence level?

Design	# Visitors	Total Transactions	St. Dev
Old	10000	1200	150
New	10000	1700	200



$$H_0 : \mu_{new} - \mu_{old} = 0$$
$$H_1 : \mu_{new} - \mu_{old} \neq 0$$

$$z = \frac{(X_{new} - X_{old}) - (\mu_{new} - \mu_{old})}{\sigma_{new-old}} = \frac{(500) - (0)}{\sqrt{150^2 + 200^2}} = 2$$

From z-table, $P(Z \leq 2) = 0.977$. Since H_1 is an inequality statement we are running a two tailed z-test. $0.977 > 0.975$, we reject H_0.

Statistical hypothesis testing

2. The χ^2 statistic
   - Testing whether a sequence of $\{1, 2, \ldots, k\}$ outcomes comes from a particular k-sided die
   - Testing the independence of two variables

Testing a k-sided die

We have used the z-statistic to:

- Test whether the mean of a distribution is a certain value.
- Test whether two distributions have the same mean.

Eg. Checking whether a coin is fair.

But what if we want to check whether a k-sided die is fair?

- Rather like checking k different means, one for each outcome.
- Or, more precisely, $k-1$ different means.
- Could run $k-1$ separate tests.

Instead: run a single combined test with the χ^2 statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{((\text{observed frequency of } i) - (\text{expected frequency of } i))^2}{(\text{expected frequency of } i)}$$

and compare it to χ^2 distribution with $k-1$ degrees of freedom.

Example: is a die fair?

A gambler is concerned that the casino's die is loaded. He observes the following frequencies in a sequence of 60 tosses:

Outcome	1	2	3	4	5	6
Observed	4	6	17	16	8	9
Expected	10	10	10	10	10	10

Null hypothesis: die is fair.

Compute the χ^2 statistic for this data:

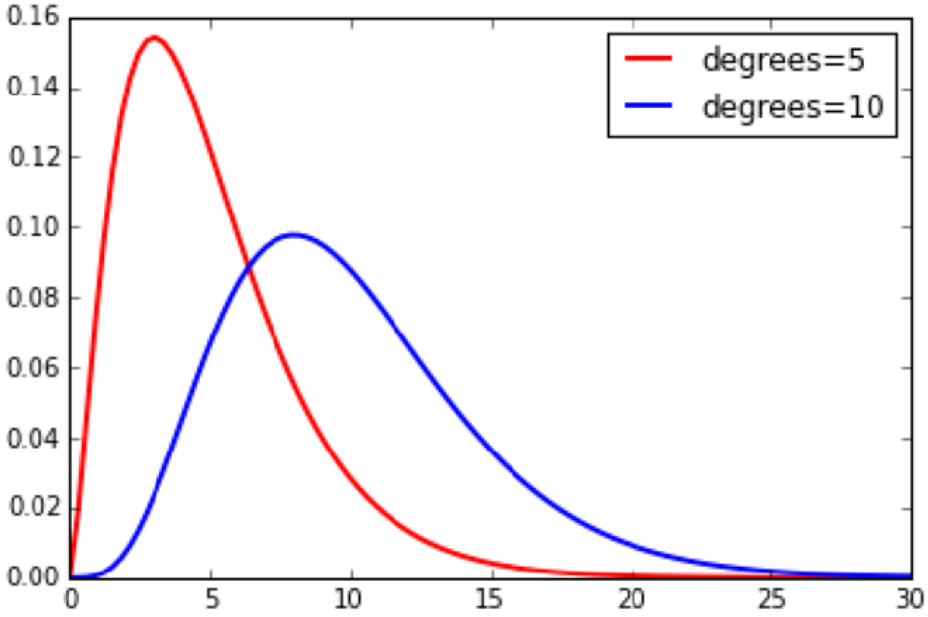
$$\chi^2 = \sum_{i=1}^{k} \frac{((\text{observed frequency of } i) - (\text{expected frequency of } i))^2}{(\text{expected frequency of } i)}$$

$$= \frac{6^2}{10} + \frac{4^2}{10} + \frac{7^2}{10} + \frac{6^2}{10} + \frac{2^2}{10} + \frac{1^2}{10} = 14.2$$

Under the null, this value would be a random draw from a χ^2 distribution with 5 degrees of freedom.

Testing fairness of a die, cont'd

The χ^2 distribution:



The probability of getting a value as large as 14.2 (with 5 degrees of freedom) is 1.4%... strong evidence against the null.

Testing independence

Suppose there are k possible outcomes.

You have two sets of observations, $S_1, S_2 \subset \{1, 2, \ldots, k\}$.
Are they independent draws from the same distribution over $\{1, \ldots, k\}$?

- Null hypothesis: They are independent.
- Estimate the underlying distribution by combining the two samples. Call this P.
- Use the χ^2 statistic of how close S_1 and S_2 are to expected frequencies under P.

Example: left-handedness by sex

Data from a sample of 2,237 Americans of age 25-34:

	Men	Women
Right-handed	934 (87.5%)	1,070 (91.5%)
Left-handed	113 (10.6%)	92 (7.9%)
Ambidextrous	20 (1.9%)	8 (0.7%)

Is left-handedness really more common in men, or is this just a chance effect from sampling?

Null hypothesis: The two sets of numbers (for men and women) are independent draws from the same distribution.

Left-handedness, cont'd

Estimate the underlying distribution as well as expected frequencies for each of the two samples:

	Observed			Expected	
	Men	Women	Total	Men	Women
Right-handed	934	1,070	2,004 (89.6%)	956	1,048
Left-handed	113	92	205 (9.2%)	98	107
Ambidextrous	20	8	28 (1.2%)	13	15
Total	1,067	1,170	2,237	1,067	1,170

Compute the χ^2 statistic for this data:

$$\chi^2 = \sum_{\text{outcomes}} \frac{((\text{observed frequency}) - (\text{expected frequency}))^2}{(\text{expected frequency})}$$

$$= \frac{22^2}{956} + \frac{22^2}{1,048} + \frac{15^2}{98} + \frac{15^2}{107} + \frac{7^2}{13} + \frac{7^2}{15} \approx 12$$

Under the null, this would have a χ^2 distribution with 2 degrees of freedom. A value ≥ 12 has probability roughly 0.2%.

Chi-Square Test: Degree of Freedom

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

- where the square of the differences between the observed and expected values in each cell, divided by the expected value, are added across all of the cells in the table.

- The distribution of the statistic X^2 is **chi-square** with (r-1)(c-1) degrees of freedom, where r represents the number of rows in the two-way table and c represents the number of columns.

- The chi-square distribution is defined for all positive values. The P-value for the chi-square test is $P(\chi^2 > X^2)$, the probability of observing a value at least as extreme as the test statistic for a chi-square distribution with (r-1)(c-1) degrees of freedom.