Introduction to Statistical Inference

Lecture 5: Survey Sampling

Mohammad-Reza A. Dehaqani

dehaqani@ut.a

Survey Sampling: What and Why

In **surveys sampling** we try to obtain information about a large population based on a relatively small sample of that population.

The main goal of **survey sampling** is to reduce the cost and the amount of work that it would take to explore the entire population.

First examples: Graunt (1662) and Laplace (1812) used survey sampling to estimate the population of London and France, respectively.

Mathematical Framework

Suppose that the target population is of size N (N is large) and a numerical value of interest x_i (age, weight, income, etc) is associated with i^{th} member of the population, $i = 1, \ldots, N$. Population parameters (quantities we are interested in):

- Population mean

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- Population variance

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

A useful identity can be obtained by expanding the square in this equation:

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - 2\mu\sum_{i=1}^{N} x_i + N\mu^2\right)\\ &= \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - 2N\mu^2 + N\mu^2\right)\\ &= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\end{aligned}$$

In the dichotomous case, the population variance reduces to $p(1-p)$:

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\\ &= p - p^2\\ &= p(1-p)\end{aligned}$$

- We are interested in population parameters:
  - Population mean $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$
  - Population variance $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$

There are several ways to sample from a population. We discussed two:

1. Simple Random Sampling

Definition

In Simple Random Sampling, each member is chosen entirely by chance and, therefore, each member has an equal chance of being included in the sample; each particular sample of size n has the same probability of occurrence.

If $X_1, \ldots, X_n$ is the sample drawn from the population, then the sample mean is a natural estimate of the population mean μ:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \approx \mu$$

2 Stratified Random Sampling

Definition

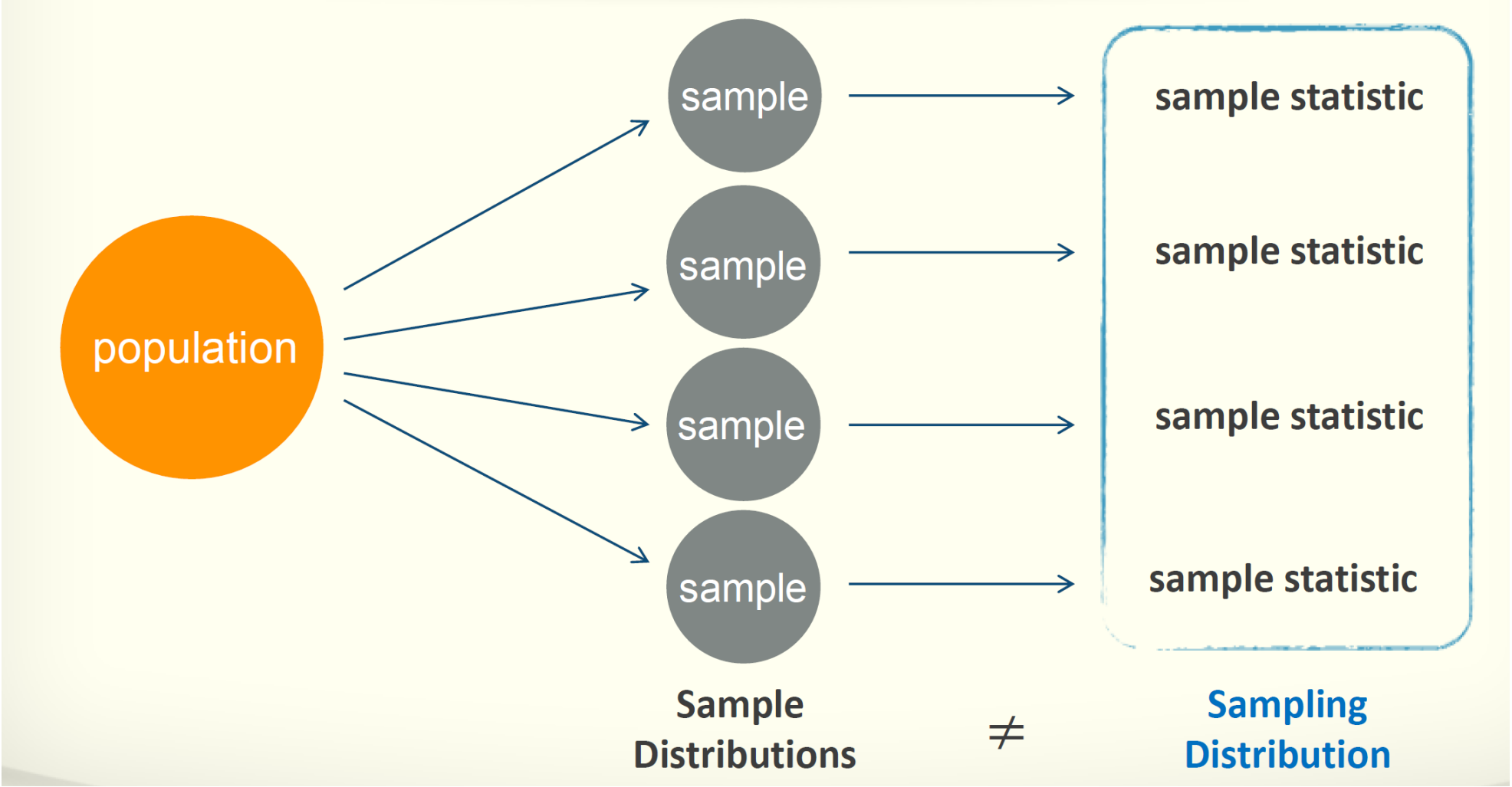
In Stratified Random Sampling, the population is partitioned into subpopulations, or strata, which are then independently sampled using simple random sampling.

If $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$ is the sample drawn from the k^{th} stratum, then the natural estimate of μ is

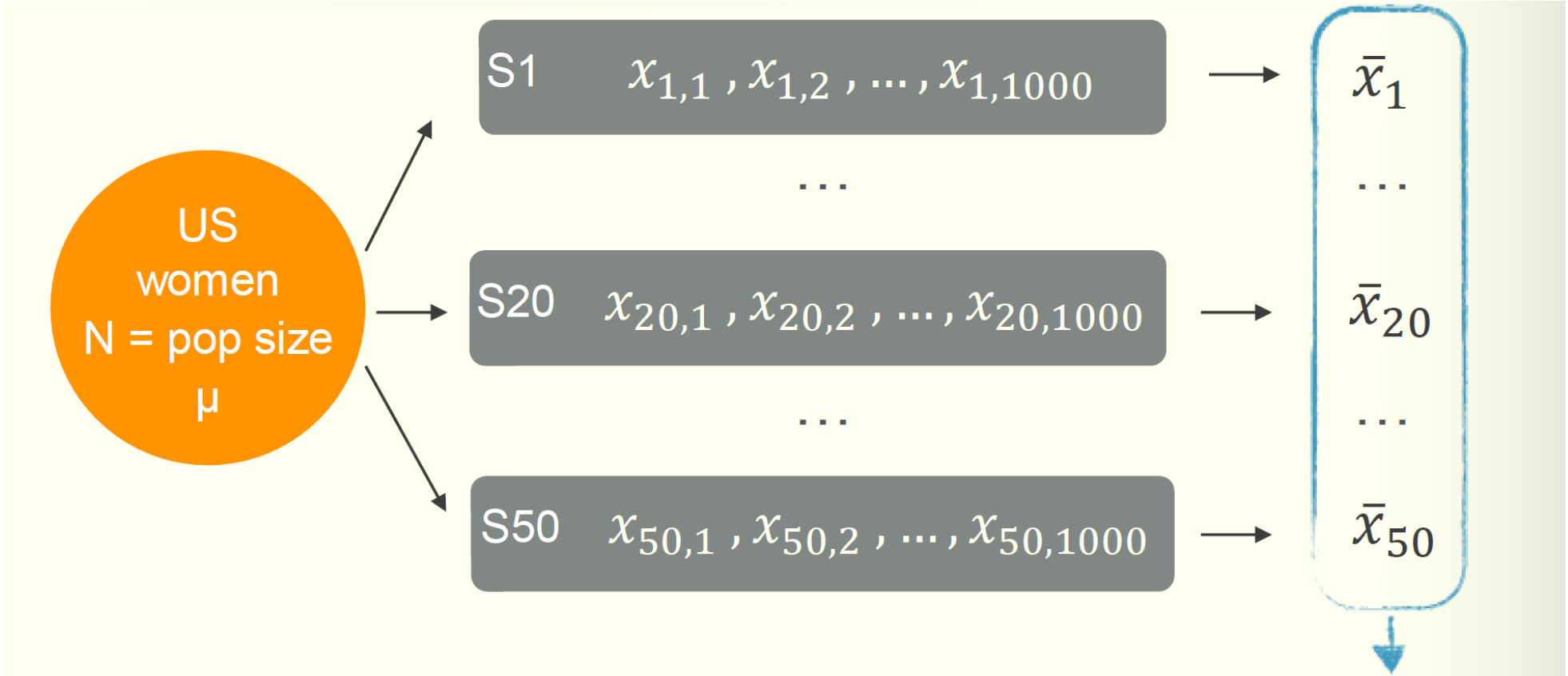
$$\overline{X}_n^* = \sum_{k=1}^{L} \omega_k \overline{X}_{n_k}^{(k)} \approx \mu$$

where ω_k is the fraction of the population in the k^{th} stratum.

Sample vs. Sampling Distribution



Example for sample mean



$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

$\bar{x}$: sampling distribution

$$mean(\bar{x}) = \mu$$

$SD(\bar{x}) < \sigma \longrightarrow$ standard error

Statistical Properties of $\overline{X}_n$

Since $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, statistical properties of $\overline{X}_n$ are completely determined by statistical properties of X_i.

Lemma

distribution of a single sample element, X_i

Denote the distinct values assumed by the population members by $\xi_1, \ldots, \xi_m$, $m \leq N$, and denote the number of population members that have the value ξ_i by n_i. Then X_i is a discrete random variable with probability mass function

$$\mathbb{P}(X_i = \xi_j) = \frac{n_j}{N}$$

Also

$$\mathbb{E}[X_i] = \mu \qquad \mathbb{V}[X_i] = \sigma^2$$

From this lemma, it follows immediately that $\overline{X}_n$ is an unbiased estimate of μ:

$$\mathbb{E}[\overline{X}_n] = \mu$$

Thus, on average $\overline{X}_n = \mu$.

The following lemma holds whether sampling is with or without replacement

$\overline{X}_n$ is an unbiased estimator of μ

This result can be interpreted as follows: "on average" $\overline{X}_n = \mu$

Definition

Suppose we want to estimate a parameter θ by a function $\hat{\theta}$ of the sample $X_1, \ldots, X_n$,

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$$

The estimator $\hat{\theta}$ is called **unbiased** if $\mathbb{E}[\hat{\theta}] = \theta$

Thus, $\overline{X}_n$ is an unbiased estimator of μ

$\text{MSE} = \beta^2 + \sigma^2.$

Recall

Proof

From Theorem B of Section 4.2,

$$\begin{aligned} E[(X - x_0)^2] &= \text{Var}(X - x_0) + [E(X - x_0)]^2 \\ &= \text{Var}(X) + \beta^2 \\ &= \sigma^2 + \beta^2 \end{aligned}$$

$$\text{Mean squared error} = \text{variance} + \text{bias}^2$$

Since $\bar{X}$ is unbiased, the mean squared errors is equal to its variance.

Statistical Properties of $\overline{X}_n$

The next important question is how variable $\overline{X}_n$ is.

As a measure of the dispersion of $\overline{X}_n$ about μ, we use the standard deviation of $\overline{X}_n$, denoted as $\sigma_{\overline{X}_n} = \sqrt{\mathbb{V}[\overline{X}_n]}$.

$$\mathbb{V}[\overline{X}_n] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right]$$

Remark: If sampling were done with replacement then X_i would be independent, and we would have:

$$\mathbb{V}[\overline{X}_n] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

In simple random sampling, we do sampling without replacement. This induces dependence among X_i. And therefore

$$\mathbb{V}[\overline{X}_n] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] \neq \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i]$$

$$\mathbb{V}\left[\sum_{i=1}^{n}\alpha_i X_i\right] = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j \mathrm{Cov}(X_i, X_j)$$

$$\mathbb{V}[\overline{X}_n] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i, X_j)$$

Lemma

If $i \neq j$, then the covariance between X_i and X_j is

$$\mathrm{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$$

$$\mathrm{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$$

$$\begin{aligned} E(X_i X_j) &= \sum_{k=1}^{m}\sum_{l=1}^{m} \zeta_k \zeta_l P(X_i = \zeta_k \text{ and } X_j = \zeta_l) \\ &= \sum_{k=1}^{m} \zeta_k P(X_i = \zeta_k) \sum_{l=1}^{m} \zeta_l P(X_j = \zeta_l | X_i = \zeta_k) \end{aligned}$$

$$P(X_j = \zeta_l | X_i = \zeta_k) = \begin{cases} n_l/(N-1), & \text{if } k \neq l \\ (n_l - 1)/(N-1), & \text{if } k = l \end{cases}$$

Theorem

The variance of $\overline{X}_n$ is given by

$$\boxed{\mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)}$$

Important observations:

- If $n << N$, then

$$\mathbb{V}[\overline{X}_n] \approx \frac{\sigma^2}{n} \qquad \sigma_{\overline{X}_n} \approx \frac{\sigma}{\sqrt{n}}$$

  $\left(1 - \frac{n-1}{N-1}\right)$ is called finite population correction. This factor arises because of dependence among X_i.

- To double the accuracy of $\mu \approx \overline{X}_n$, the sample size must be quadrupled

- If σ is small (the population values are not very dispersed), then a small sample will be fairly accurate. But if σ is large, then a larger sample will be required to obtain the same accuracy.

Estimation of the Population Variance σ

- Why do we need to estimate σ?

$$\sigma_{\overline{X}_n} = \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)}, \qquad \sigma_{\overline{X}_n} \approx \frac{\sigma}{\sqrt{n}}, \quad \text{if } n \ll N \qquad (1)$$

We can't use (1) since σ is unknown.

To use (1), σ must be estimated from the sample $X_1, \ldots, X_n$.

Estimation of σ

It seems natural to use the following estimate

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

However, this estimate is biased.

Theorem

The expected value of $\hat{\sigma}_n^2$ is given by

$$\mathbb{E}[\hat{\sigma}_n^2] = \sigma^2 \frac{Nn - N}{Nn - n}$$

$$E(\hat{\sigma}^2) = \sigma^2 \left(\frac{n-1}{n}\right) \frac{N}{N-1}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 \quad \Longrightarrow \quad E(\hat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i^2\right) - E(\overline{X}^2)$$

$$\begin{aligned} E\left(X_i^2\right) &= \mathrm{Var}(X_i) + [E(X_i)]^2 \\ &= \sigma^2 + \mu^2 \end{aligned}$$

$$\begin{aligned} E(\overline{X}^2) &= \mathrm{Var}(\overline{X}) + [E(\overline{X})]^2 \\ &= \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) + \mu^2 \end{aligned}$$

$$\mathbb{E}[\hat{\sigma}_n^2] = \sigma^2 \frac{Nn - N}{Nn - n}$$

Important Remark:

- Since $\frac{Nn-N}{Nn-n} < 1$, we have $\mathbb{E}[\hat{\sigma}_n^2] < \sigma^2$
  Therefore, $\hat{\sigma}_n^2$ tends to underestimate σ^2

Since $\mathbb{E}[\hat{\sigma}_n^2] = \sigma^2 \frac{Nn-N}{Nn-n}$,

$$\hat{\sigma}^2_{n,\text{unbiased}} = \frac{Nn-n}{Nn-N}\hat{\sigma}_n^2$$

is an unbiased estimate of σ^2

Recall that

$$\mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$

An unbiased estimate of $\mathbb{V}[\overline{X}_n]$ *is*

$$\boxed{s^2_{\overline{X}_n} = \frac{\hat{\sigma}_n^2}{n}\frac{Nn-n}{Nn-N}\left(1 - \frac{n-1}{N-1}\right)}$$

Summary

Let us summarize what we have learned about estimation of population parameters:

- **Population mean** μ
  - Unbiased estimate:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  - Variance of estimate

$$\mathbb{V}[\overline{X}_n] \equiv \sigma^2_{\overline{X}_n} = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$

  - Estimated variance

$$\sigma^2_{\overline{X}_n} \approx s^2_{\overline{X}_n} = \frac{\hat{\sigma}^2_n}{n}\frac{Nn-n}{Nn-N}\left(1 - \frac{n-1}{N-1}\right)$$

- **Population variance** σ
  - Unbiased estimate:

$$\hat{\sigma}^2_{n,\text{unbiased}} = \frac{Nn-n}{Nn-N}\hat{\sigma}^2_n, \qquad \hat{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Conclusion

In simple random sampling, we can not only form estimate of unknown population parameter (e.g. μ), but also obtain the likely size of errors of these estimates.In other words, we can obtain the estimate of a parameter as well as the estimate of the error of that estimate

We previous Lectures, we found the mean and the variance of the sample mean:

$$\mathbb{E}[\overline{X}_n] = \mu \qquad \mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$

Ideally, we would like to know the **entire distribution** of $\overline{X}_n$ (sampling distribution) since it would tell us everything about the random variable $\overline{X}_n$

Reminder:

If $X_1, \ldots, X_n$ are i.i.d. with the common mean μ and variance σ^2, then the sample mean $\overline{X}_n$ has the following properties:

1. $\mathbb{E}[\overline{X}_n] = \mu, \qquad \mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}$
2. CLT:

$$\mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) \to \Phi(z), \qquad \text{as } n \to \infty$$

where $\Phi(z)$ is the CDF of $\mathcal{N}(0,1)$

Q: Can we use these results to obtain the distribution of $\overline{X}_n$?

A: No. In simple random sampling, X_i are not independent.
Moreover, it makes no sense to have n tend to infinity while N is fixed.

Nevertheless, it can be shown that if n is large, but still small relative to N, then $\overline{X}_n$ is **approximately normally distributed**

$$\boxed{\overline{X}_n \dot{\sim} \mathcal{N}(\mu, \sigma^2_{\overline{X}_n})} \qquad \sigma_{\overline{X}_n} = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n-1}{N-1}}$$

Suppose we want to find the probability that the error made in estimating μ by $\overline{X}_n$ is less than $\varepsilon > 0$. In symbols, we want to find

$$\mathbb{P}(|\overline{X}_n - \mu| \leq \varepsilon) = ?$$

Theorem

From $\overline{X}_n \dot{\sim} \mathcal{N}(\mu, \sigma^2_{\overline{X}_n})$ it follows that

$$\boxed{\mathbb{P}(|\overline{X}_n - \mu| \leq \varepsilon) \approx 2\Phi\left(\frac{\varepsilon}{\sigma_{\overline{X}_n}}\right) - 1}$$

Confidence Intervals

Let $\alpha \in [0, 1]$

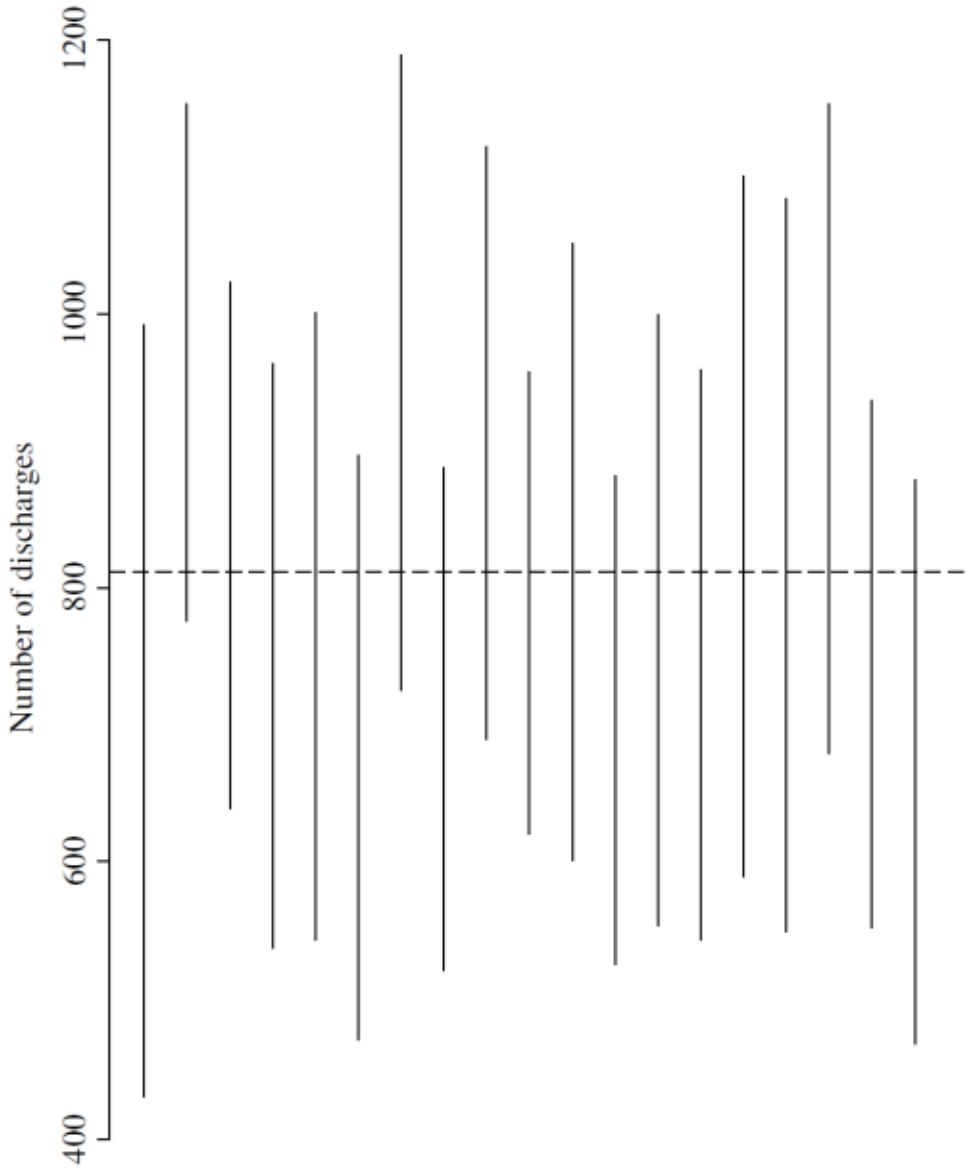
Definition

A $100(1-\alpha)\%$ **confidence interval** for a population parameter θ is a random interval calculated from the sample, which contains θ with probability $1-\alpha$.

Interpretation:

If we were to take many random samples and construct a confidence interval from each sample, then about $100(1-\alpha)\%$ of these intervals would contain θ.

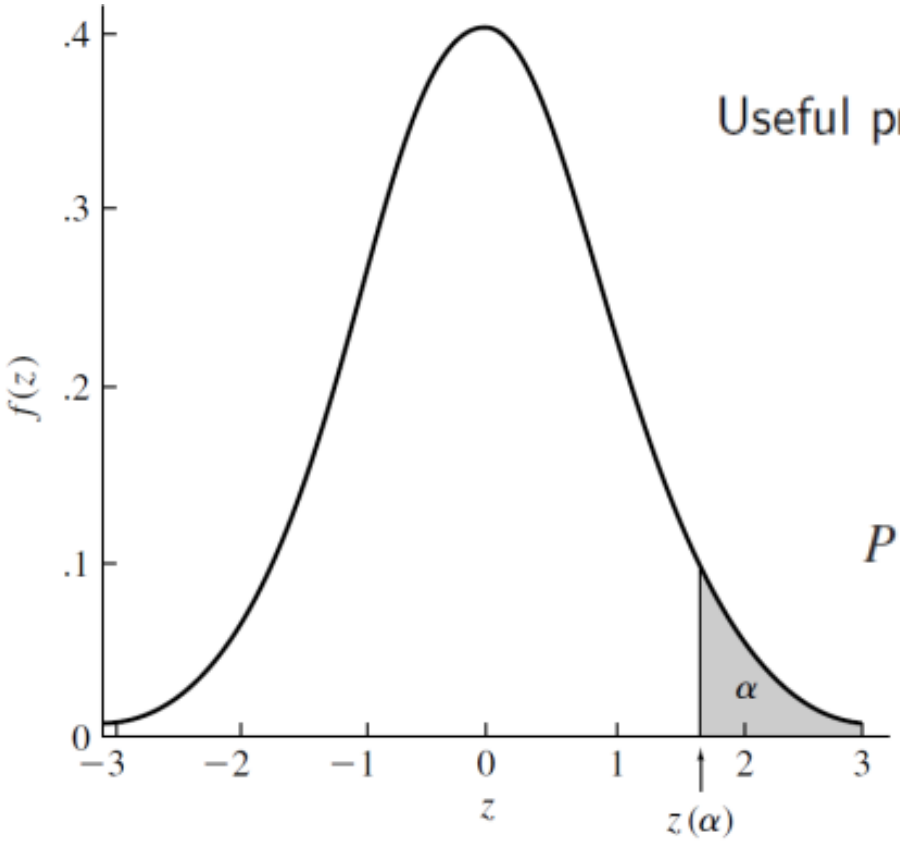
20 samples each of size $n = 25$



Our goal: to construct a confidence interval for μ

Let z_α be that number such that the area under the standard normal density function to the right of z_α is α. In symbols, z_α is such that

$$\Phi(z_\alpha) = 1 - \alpha$$

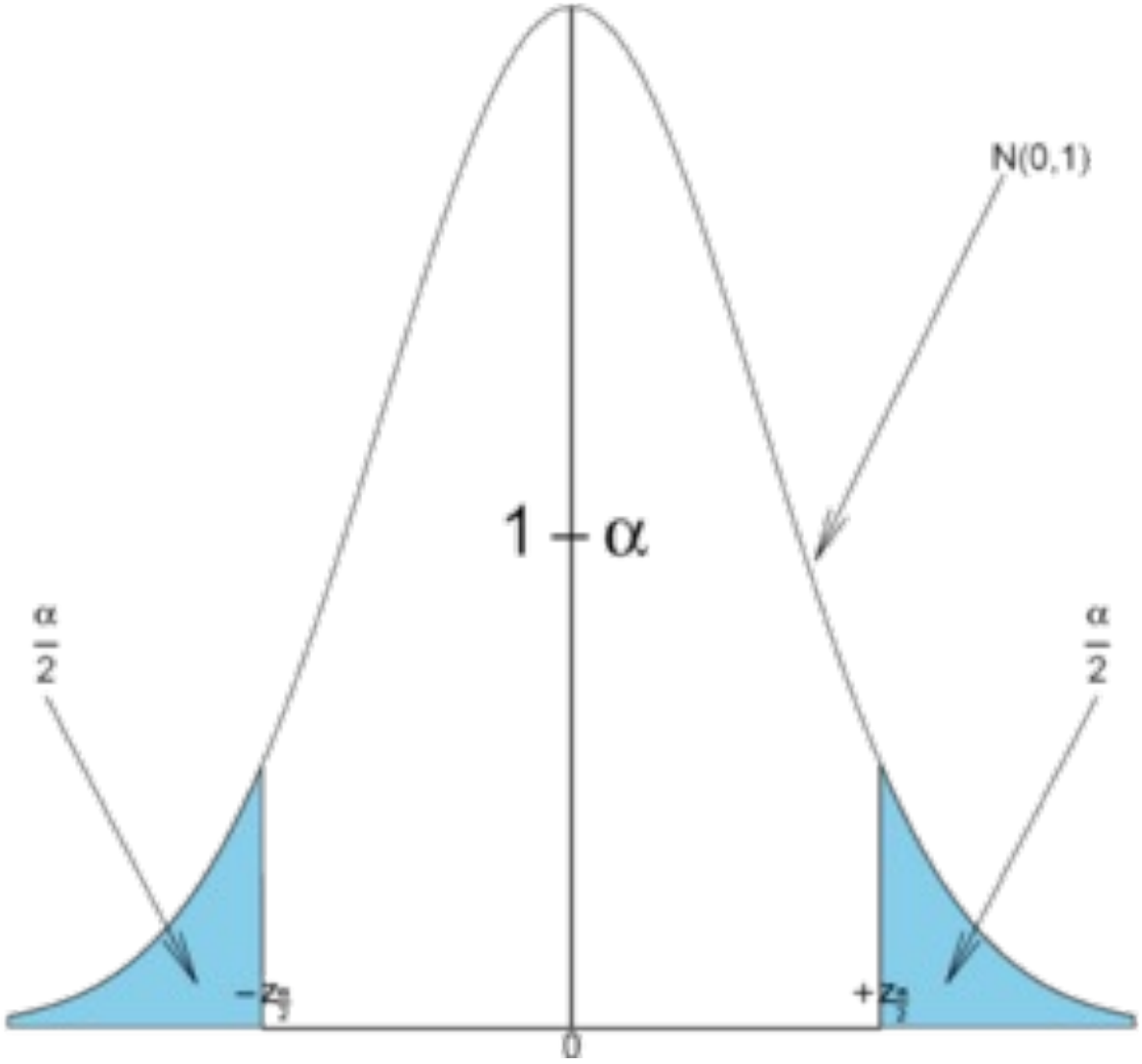


Useful property:

$$z_{1-\alpha} = -z_\alpha$$

$$P\left(-z(\alpha/2) \leq \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \leq z(\alpha/2)\right) \approx 1 - \alpha$$

$$P\left(-z(\alpha/2) \leq \frac{\overline{X}-\mu}{\sigma_{\overline{X}}} \leq z(\alpha/2)\right) \approx 1-\alpha$$



Confidence interval for μ

Theorem

An (approximate) $100(1-\alpha)\%$ *confidence interval for* μ *is*

$$(\overline{X}_n - z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}, \overline{X}_n + z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n})$$

That is the probability that μ *lies in that interval is approximately* $1-\alpha$

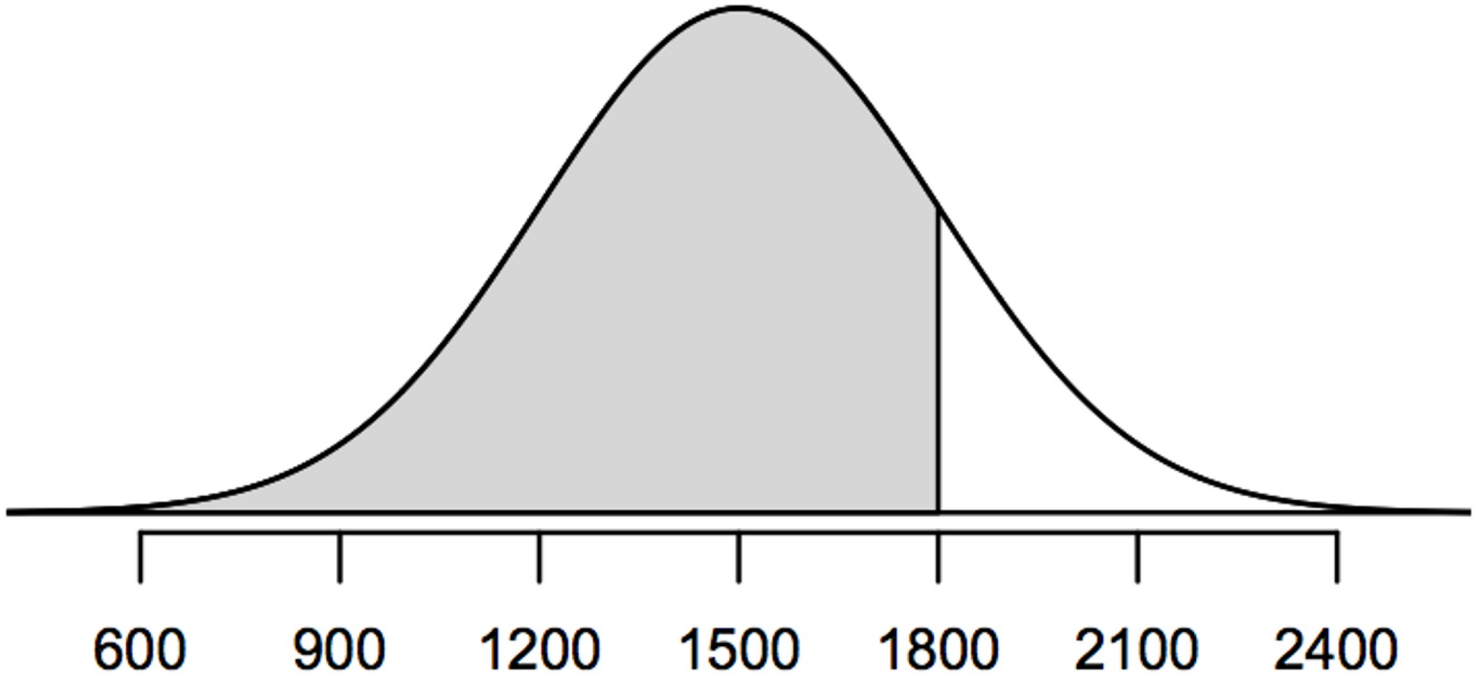
$$\boxed{\mathbb{P}(\overline{X}_n - z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n} \leq \mu \leq \overline{X}_n + z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}) \approx 1-\alpha}$$

Remarks:

- This confidence interval is random. The probability that it covers μ is $(1-\alpha)$
- In practice, $\alpha = 0.1, 0.05, 0.01$ (depends on a particular application)
- Since $\sigma_{\overline{X}_n}$ is not known (it depends on σ), $s_{\overline{X}_n}$ is used instead of $\sigma_{\overline{X}_n}$

Percentiles

- *Percentile* is the percentage of observations that fall below a given data point.
- Graphically, percentile is the area below the probability distribution curve to the left of that observation.

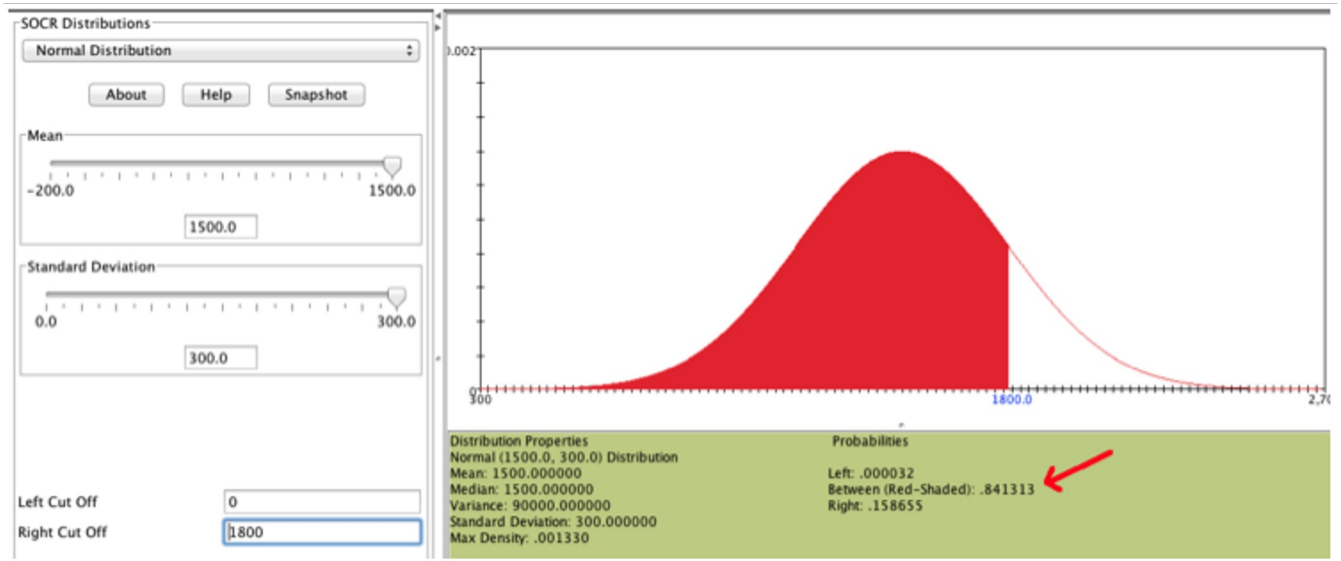


Calculating percentiles - using computation

There are many ways to compute percentiles/areas under the curve. R:

```
> pnorm(1800, mean = 1500, sd = 300)
[1] 0.8413447
```

Applet: www.socr.ucla.edu/htmls/SOCR_Distributions.html



Calculating percentiles - using tables

	Second decimal place of Z									
Z	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

Six sigma

The term *six sigma process* comes from the notion that if one has six standard deviations between the process mean and the nearest specification limit, as shown in the graph, practically no items will fail to meet specifications.

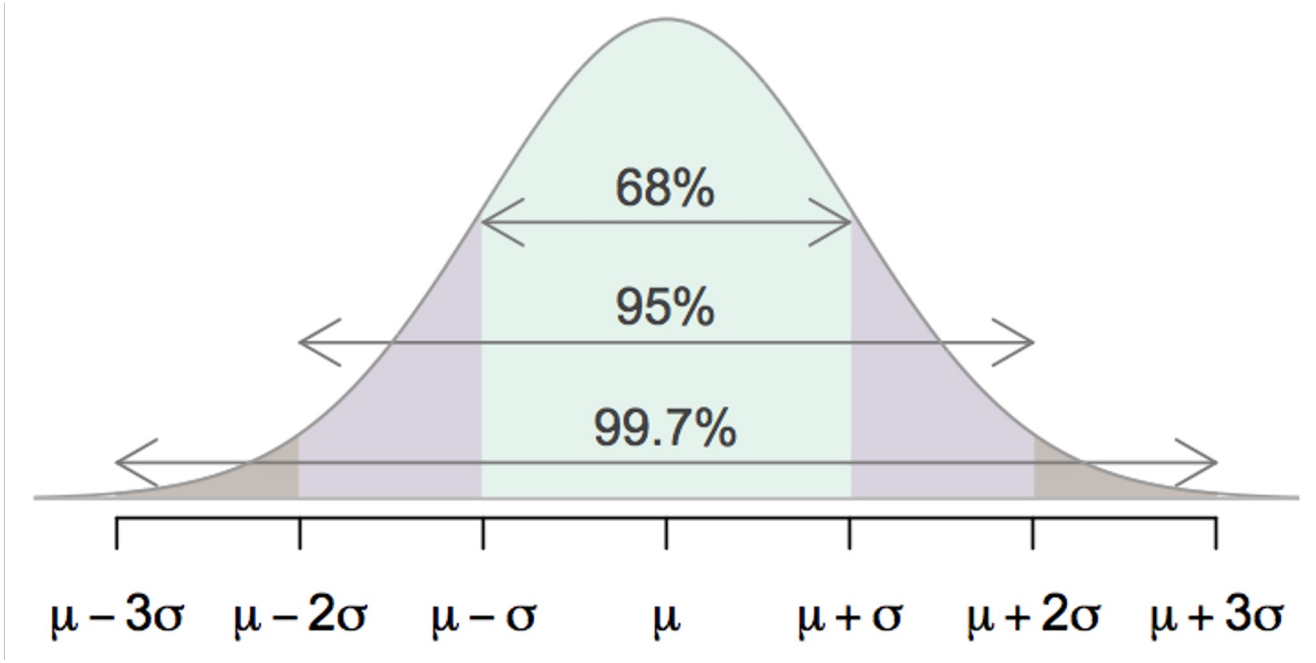
http://en.wikipedia.org/wiki/Six_Sigma

68-95-99.7 Rule

For nearly normally distributed data,

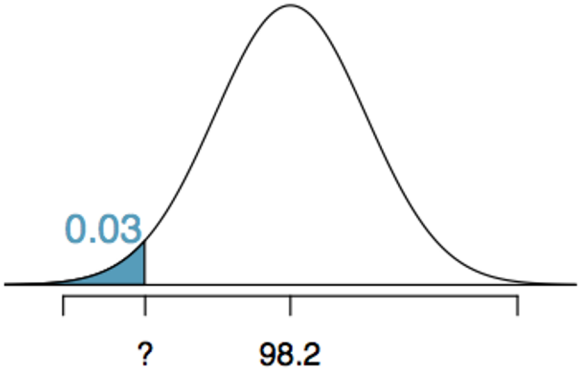
- about 68% falls within 1 SD of the mean,
- about 95% falls within 2 SD of the mean,
- about 99.7% falls within 3 SD of the mean.

It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



Finding cutoff points

Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the lowest 3% of human body temperatures?



0.09	0.08	0.07	0.06	0.05	Z
0.0233	0.0239	0.0244	0.0250	0.0256	-1.9
0.0294	*0.0301*	0.0307	0.0314	0.0322	-1.8
0.0367	0.0375	0.0384	0.0392	0.0401	-1.7

$$P(X < x) = 0.03 \rightarrow P(Z < -1.88) = 0.03$$

$$Z = \frac{obs - mean}{SD} \rightarrow \frac{x - 98.2}{0.73} = -1.88$$

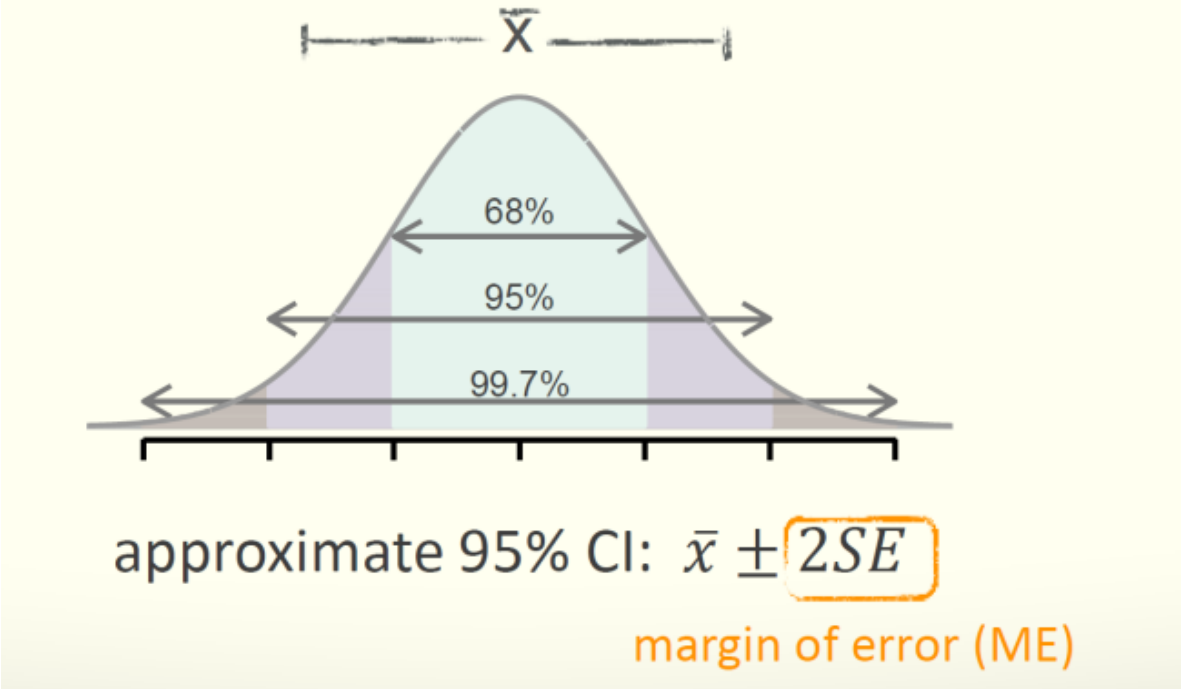
$$x = (-1.88 \times 0.73) + 98.2 = 96.8°F$$

Mackowiak, Wasserman, and Levine (1992), A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlick.

Base on CLT, the distribution of sample mean approximated by a normal model:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right),$$

where ***SE*** is represents *standard error*, which is defined as the standard deviation of the sampling distribution. If σ is unknown, use s.



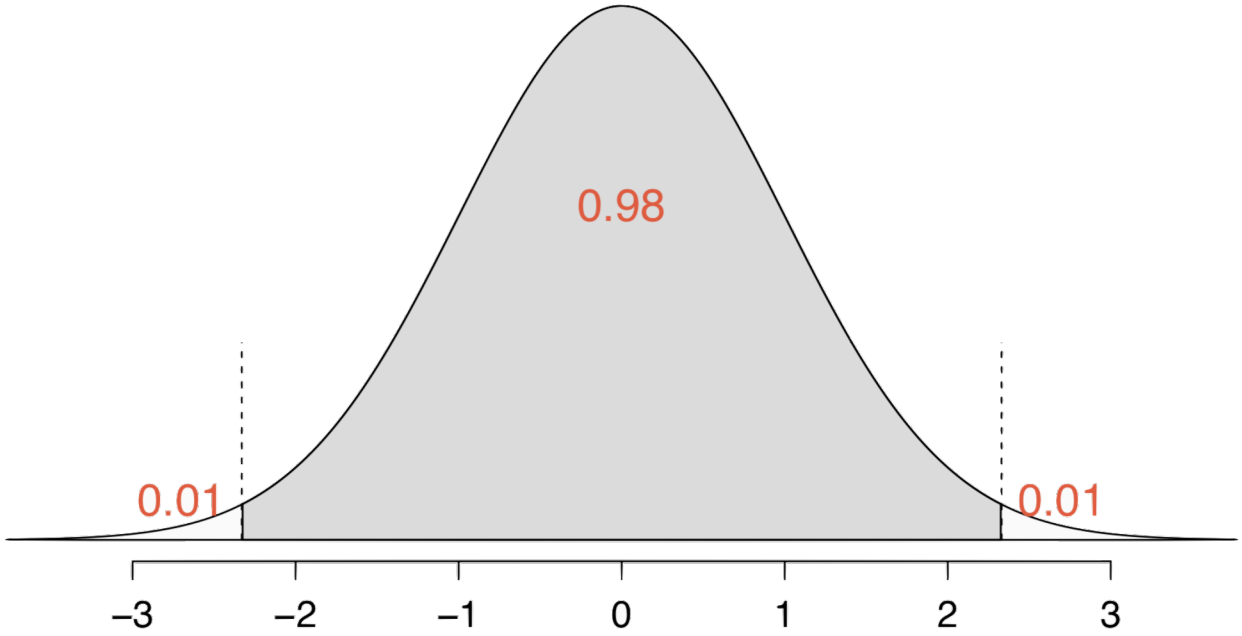
Changing the confidence level for sample mean; Margin of error

$$\text{point estimate} \pm z^\star \times SE$$

- In a confidence interval, **$z^\star \times SE$** is called the **margin of error**, and for a given sample, the margin of error changes as the confidence level changes.
- For a 95% confidence interval, $z^\star = 1.96$.
- However, using the standard normal (z) distribution, it is possible to find the appropriate $z^\star$ for any confidence level.

Example:

Appropriate $z^\star$ when calculating a 98% confidence interval



Summary

- The sample mean is approximately normal

$$\boxed{\overline{X}_n \dot{\sim} \mathcal{N}(\mu, \sigma^2_{\overline{X}_n})} \qquad \sigma_{\overline{X}_n} = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}}$$

- Probability of error

$$\mathbb{P}(|\overline{X}_n - \mu| \leq \varepsilon) \approx 2\Phi\left(\frac{\varepsilon}{\sigma_{\overline{X}_n}}\right) - 1$$

- $100(1-\alpha)\%$ confidence interval for μ is

$$(\overline{X}_n - z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}, \overline{X}_n + z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n})$$

The Sample Mean and the Sample Variance Under Assumption of Normality

Framework

Let $X_1, \ldots, X_n$ be a sample drawn from a population.

Suppose that the population is "Gaussian" $\boxed{X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)}$

We want to estimate population parameters μ and σ^2.

Definition

- The **sample mean** is $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
- The **sample variance** is $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$

Theorem

$\overline{X}_n$ and S_n^2 are **unbiased estimators** *of μ and σ^2, respectively,*

$$\mathbb{E}[\overline{X}_n] = \mu, \qquad \mathbb{E}[S_n^2] = \sigma^2$$

Our goal: to describe distributions of $\overline{X}_n$ and S_n^2

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

It is a biased estimator

It can be shown that

$$\mathrm{E}\left[\hat{\sigma}^2\right] = \frac{n-1}{n}\sigma^2$$

Proof of bias

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

$$\mathrm{E}[S^2] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left((X_i - \mu) - (\overline{X} - \mu)\right)^2\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left((X_i - \mu)^2 - 2(\overline{X} - \mu)(X_i - \mu) + (\overline{X} - \mu)^2\right)\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + \frac{1}{n}(\overline{X} - \mu)^2 \sum_{i=1}^{n} 1\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + \frac{1}{n}(\overline{X} - \mu)^2 \cdot n\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + (\overline{X} - \mu)^2\right]$$

To continue, we note that by subtracting μ from both sides of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, we get

$$\overline{X} - \mu = \frac{1}{n}\sum_{i=1}^{n} X_i - \mu = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu).$$

$$\begin{aligned}
\operatorname{E}[S^2] &= \operatorname{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + (\overline{X} - \mu)^2\right] \\
&= \operatorname{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}(\overline{X} - \mu)\cdot n \cdot (\overline{X} - \mu) + (\overline{X} - \mu)^2\right] \\
&= \operatorname{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X} - \mu)^2 + (\overline{X} - \mu)^2\right] \\
&= \operatorname{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X} - \mu)^2\right] \\
&= \operatorname{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right] - \operatorname{E}\left[(\overline{X} - \mu)^2\right] \\
&= \sigma^2 - \operatorname{E}\left[(\overline{X} - \mu)^2\right] = \left(1 - \frac{1}{n}\right)\sigma^2 < \sigma^2.
\end{aligned}$$

Distribution of $\overline{X}_n$

Theorem

If $X_1, \ldots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$ random variables, then

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

Distribution of S_n^2

Theorem

If $X_1, \ldots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$ random variables, then

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

The χ^2-distribution

Definition

Let $Z_1, \ldots, Z_n$ be independent standard normal variables,

$$Z_1, \ldots, Z_n \sim N(0, 1)$$

Then the distribution of

$$Q = Z_1^2 + Z_2^2 + \ldots + Z_n^2$$

is called the χ^2-**distribution** with n **degrees of freedom**,

$$Q \sim \chi_n^2$$

- Probability Density Function:

$$\pi(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}$$

  - $x \geq 0$
  - Γ is the gamma function $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$

Its moment-generating function is

$$M(t) = (1 - 2t)^{-n/2}$$

THEOREM **A**

The random variable $\overline{X}$ and the vector of random variables $(X_1 - \overline{X}, X_2 - \overline{X}, \ldots, X_n - \overline{X})$ are independent.

COROLLARY **A**

$\overline{X}$ and S^2 are independently distributed.

Proof

This follows immediately since S^2 is a function of the vector $(X_1 - \overline{X}, \ldots, X_n - \overline{X})$, which is independent of $\overline{X}$. ■

THEOREM **B**

The distribution of $(n-1)S^2/\sigma^2$ is the chi-square distribution with $n-1$ degrees of freedom.

Proof

We first note that

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi_n^2$$

Also,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}[(X_i-\overline{X})+(\overline{X}-\mu)]^2$$

Expanding the square and using the fact that $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$, we obtain

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 + \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2$$

This is a relation of the form $W = U + V$. Since U and V are independent by Corollary A, $M_W(t) = M_U(t)M_V(t)$. W and V both follow chi-square distributions, so

$$\begin{aligned} M_U(t) &= \frac{M_W(t)}{M_V(t)} \\ &= \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} \\ &= (1-2t)^{-(n-1)/2} \end{aligned}$$

The last expression is the mgf of a random variable with a χ^2_{n-1} distribution. ■

Nice Properties:

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

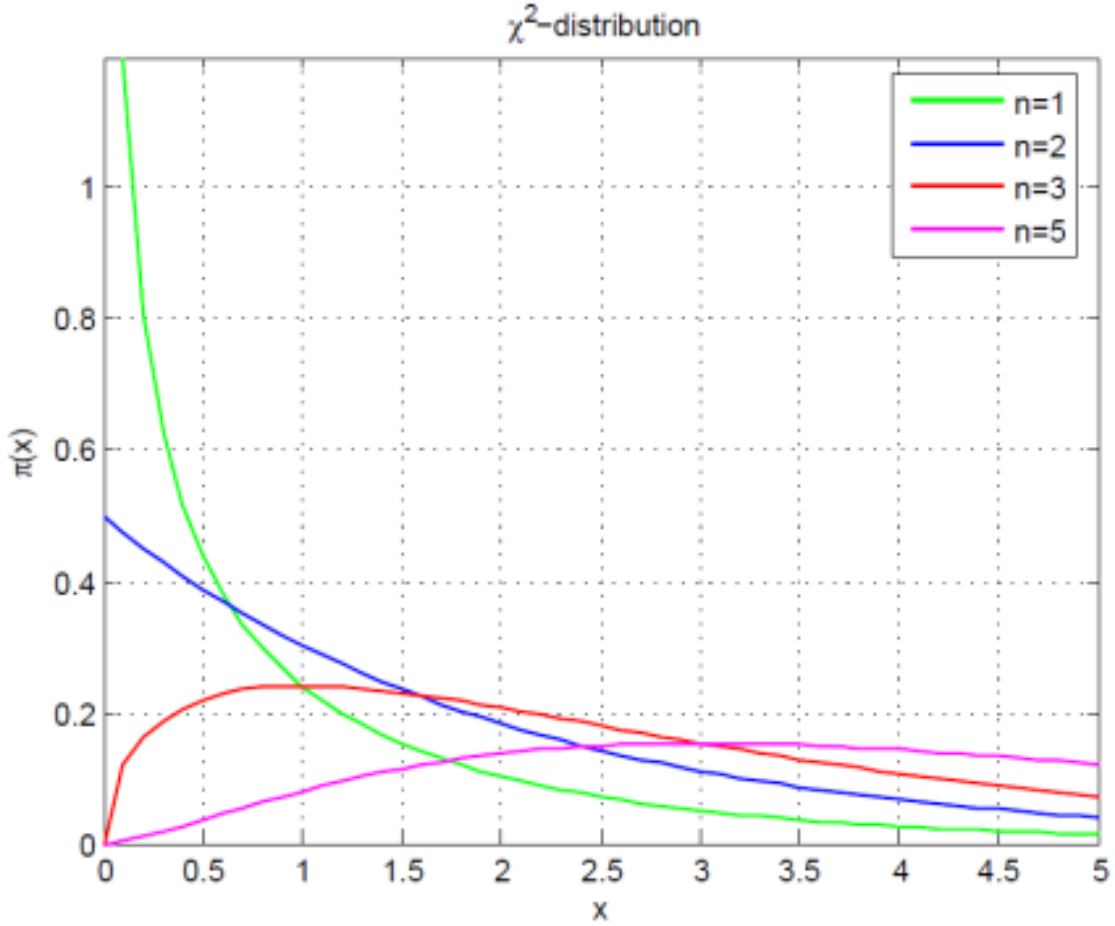
and

$$\left(\frac{X - \mu}{\sigma}\right)^2 \sim \chi^2_1$$

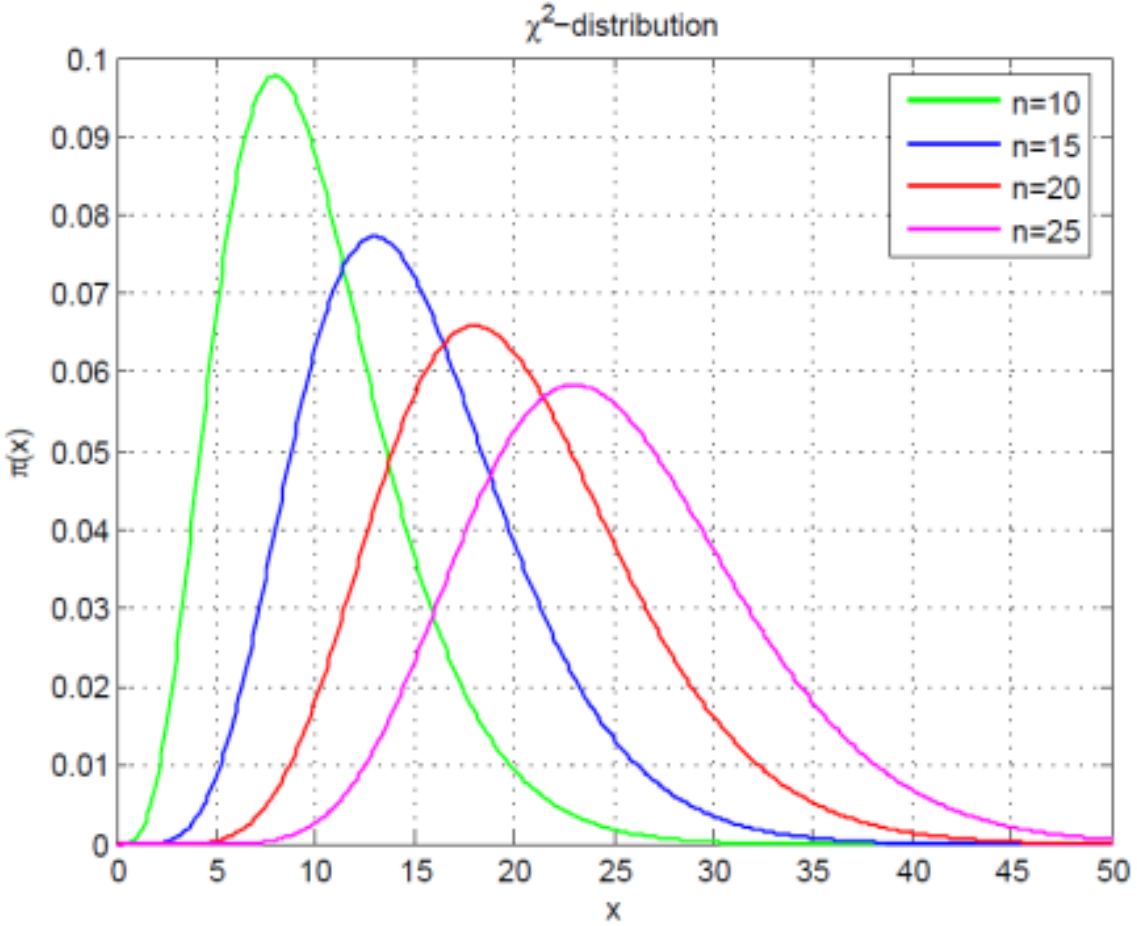
- If $U \sim \chi^2_n$ and $V \sim \chi^2_m$, and U and V are independent, then

$$U + V \sim \chi^2_{n+m}$$

Graph of the χ^2_n PDF: small n



Graph of the χ^2_n PDF: large n



- CLT: χ^2_n converges to a normal distribution as $n \to \infty$
- $\chi^2_n \to \mathcal{N}(n, 2n)$, as $n \to \infty$
- When $n > 50$, for many practical purposes, $\chi^2_n = \mathcal{N}(n, 2n)$

Bringing the t-distribution into the Game

Theorem

If $X_1, \ldots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$ random variables, then

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$$

The t-distribution

Definition

Let $Z \sim \mathcal{N}(0,1)$, $U \sim \chi^2_n$, and Z and U are independent. Then the distribution of

$$T = \frac{Z}{\sqrt{U/n}}$$

is called the t-**distribution** with n **degrees of freedom**.

- The t-distribution is symmetric about zero, $\pi(x) = \pi(-x)$
- As $n \to \infty$, the t-distribution tends to the standard normal distribution. In fact, when $n > 30$, the two distributions are very close.

Let $\overline{X}$ and S^2 be as given at the beginning of this section. Then

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

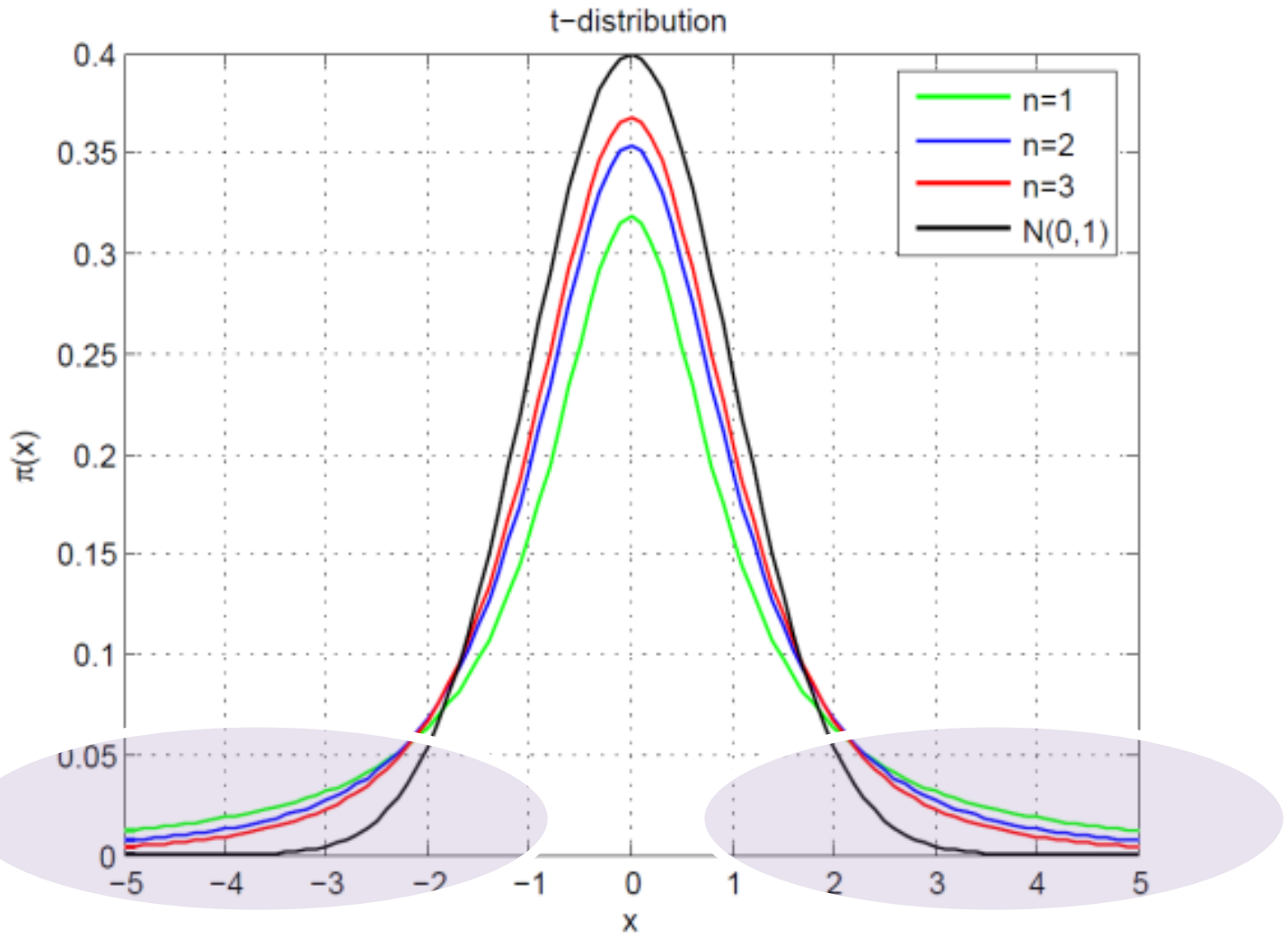
Proof

We simply express the given ratio in a different form:

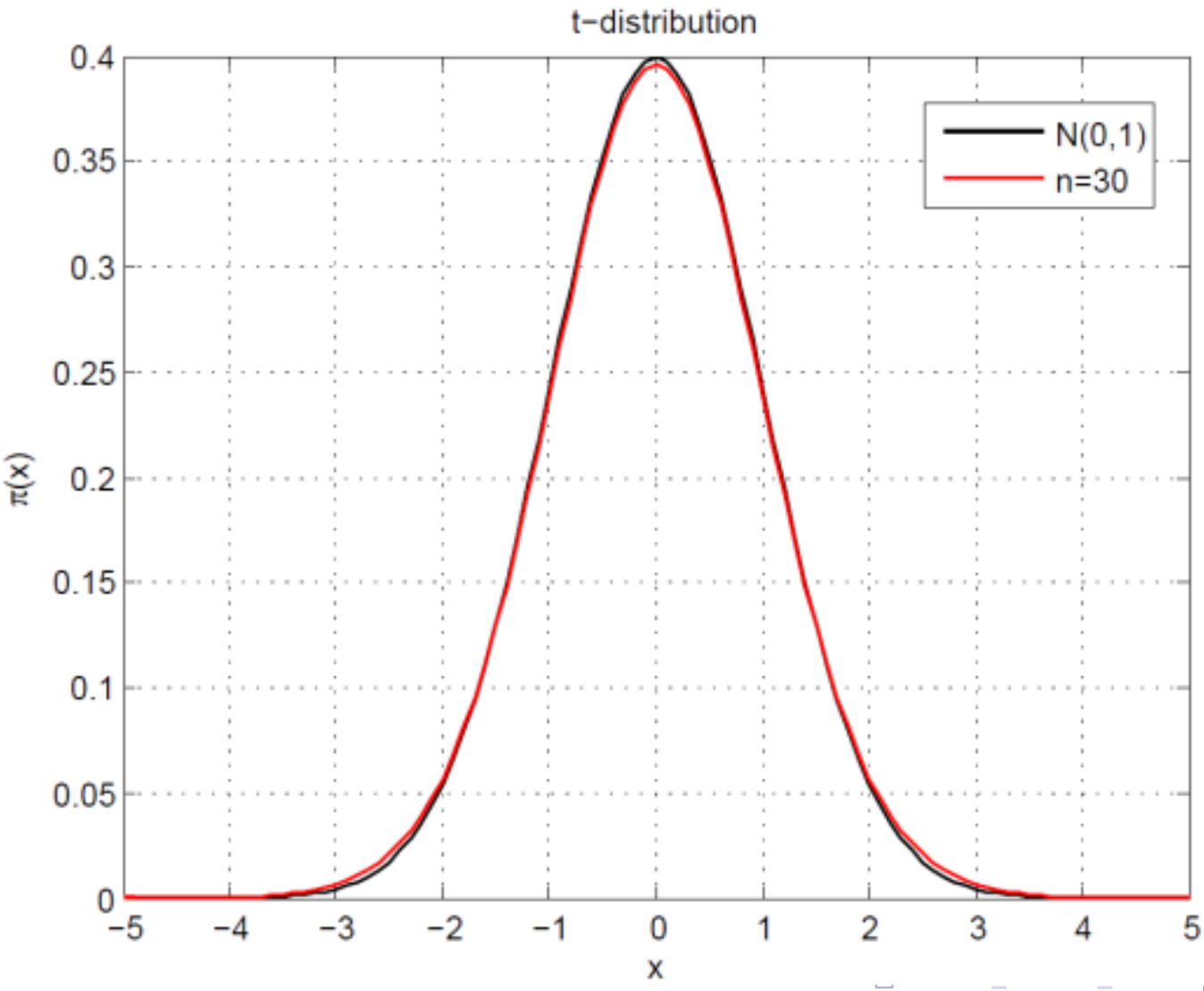
$$\frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\left(\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)}{\sqrt{S^2/\sigma^2}}$$

The latter is the ratio of an $N(0, 1)$ random variable to the square root of an independent random variable with a χ^2_{n-1} distribution divided by its degrees of freedom. Thus, from the definition in Section 6.2, the ratio follows a t distribution with $n - 1$ degrees of freedom. ■

Graph of the t-distribution PDF: small n



Graph of the t-distribution PDF: large n



F Distribution

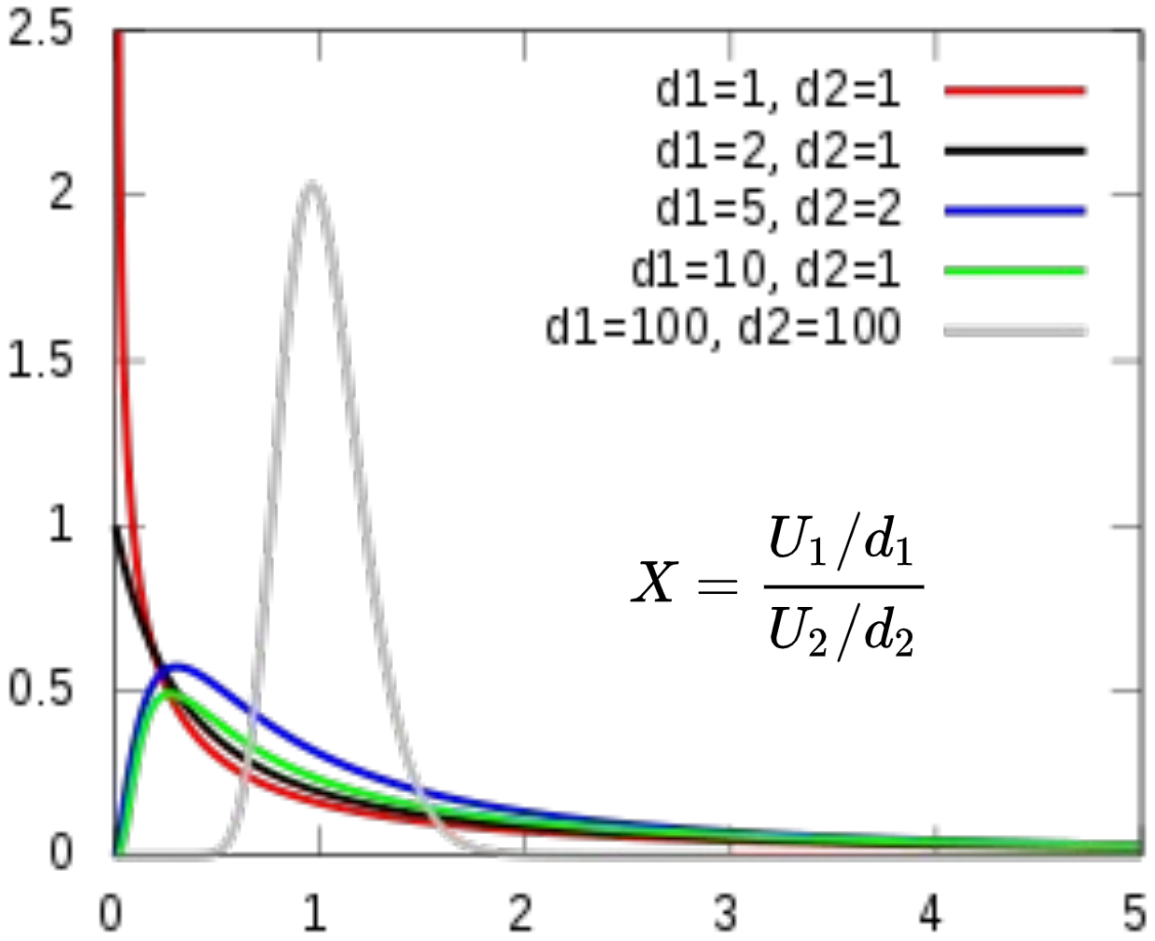
Let U and V be independent chi-square random variables with m and n degrees of freedom, respectively. The distribution of

$$W = \frac{U/m}{V/n}$$

is called the F distribution with m and n degrees of freedom and is denoted by $F_{m,n}$. ■

$$\mathbf{x} \text{ is } F(k, r) \qquad \text{then} \qquad \frac{1}{\mathbf{x}} \text{ is } F(r, k)$$

It can be shown that, for $n > 2$, $E(W)$ exists and equals $n/(n-2)$. From the definitions of the t and F distributions, it follows that the square of a t_n random variable follows an $F_{1,n}$ distribution


2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
d1=1, d2=1
d1=2, d2=1
d1=5, d2=2
d1=10, d2=1
d1=100, d2=100
$X = \frac{U_1/d_1}{U_2/d_2}$

Summary

Under Assumption of Normality, $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$,

$$\text{the sample mean:} \quad \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\text{the sample variance:} \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

have the following properties:

- $\boxed{\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)}$
- $\boxed{\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}}$ $\qquad \chi^2_n = \mathcal{N}(0,1)^2 + \ldots + \mathcal{N}(0,1)^2$
- $\boxed{\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}}$ $\qquad t_n = \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_n/n}}$

Reference

The slides contents come from USC mathematical statistics and Stanford course + John A. Rice's book + Papoulis's Book + some slides adapted from OpenIntro and Dr. Bahrak course