Chapter 5: Estimation

(Ott & Longnecker Sections: 4.12, 4.14 and 5.2)

https://dzwang91.github.io/stat324/

Part 3

WISCONSIN

UNIVERSITY OF WISCONSIN–MADISON

Outline

1. QQ plot
2. Central limit theorem
3. Review of point estimation
4. The t-distribution
5. Confidence interval

QQ plot

- We usually assume that the sample is from Normal distribution, how can we test this assumption?

QQ plot

- We usually assume that the sample is from Normal distribution, how can we test this assumption?
- One easy way is using a normal quantile-quantile plot or normal QQ plot.
- If a set of observations is approximately normally distributed, a QQ plot will result in an approximately straight line.
- R function: qqnorm(data)

QQ plot

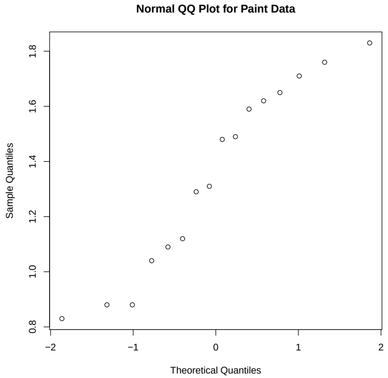
- Recall the $p \times 100\%$ quantile point q for random variable X is a point that satisfies

$$\mathbb{P}(X \leq q) = F_X(q) = p$$

  where F_X is the cumulative distribution function of X.
- The quantile q is given by $q = F_X^{-1}(p)$.
- QQ-plot of two random variables X and Y is defined to be a parametric curve $C(p)$ parameterized by $p \in [0, 1]$:

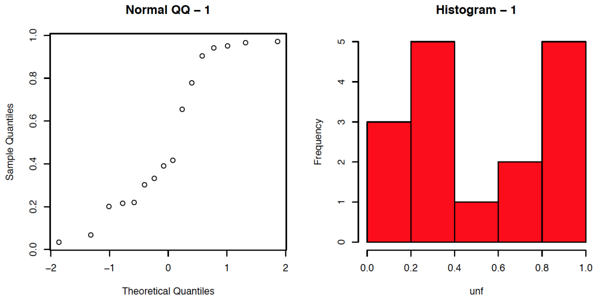
$$C(p) = (F_X^{-1}(p), F_Y^{-1}(p))$$

QQ plot example

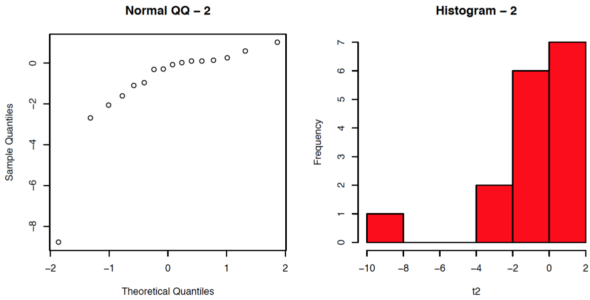


The plot is not perfectly straight, but it is pretty good.

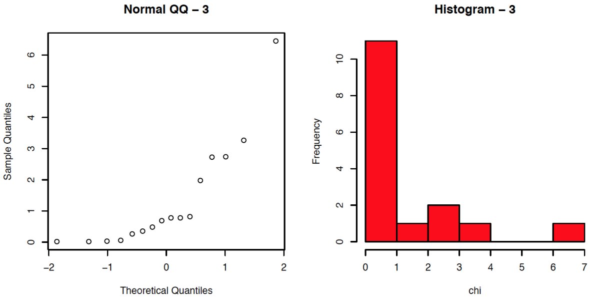
QQ plot example



QQ plot example



QQ plot example



Outline

Central limit theorem

- In many common situations, it is reasonable to assume that our sample is from a normal distribution.
  $\rightarrow$ The sample mean is also normal distributed.
- But as you can see from the QQ plot, some samples are not from Normal distribution.
  Question: what is the distribution of sample mean?

Central limit theorem

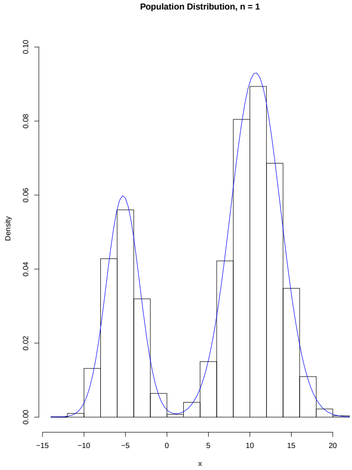
- In many common situations, it is reasonable to assume that our sample is from a normal distribution.
  $\rightarrow$ The sample mean is also normal distributed.
- But as you can see from the QQ plot, some samples are not from Normal distribution.
  Question: what is the distribution of sample mean?

Theorem

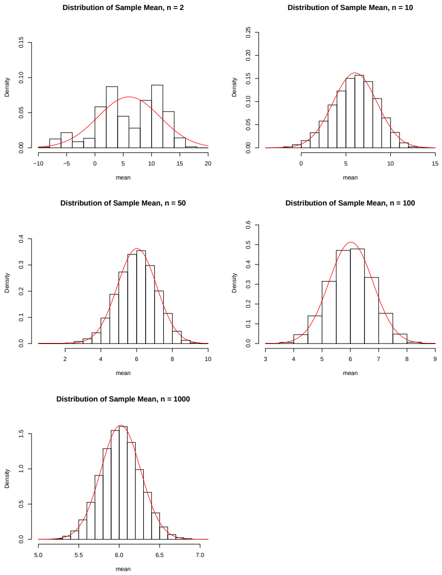
(Informal) It does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the distribution of sample means tend to follow the normal distribution as sample size is larger and larger.

Simulation

We consider the population distribution which is a mixture of two normal distributions.



Simulation


Distribution of Sample Mean, n = 2
Density
mean
Distribution of Sample Mean, n = 10
Density
mean
Distribution of Sample Mean, n = 50
Density
mean
Distribution of Sample Mean, n = 100
Density
mean
Distribution of Sample Mean, n = 1000
Density
mean

Central limit theorem

Theorem

(Formal) Let X_1, X_2, ..., X_n *be a collection of iid RVs with* $E(X_i) = \mu$ *and* $VAR(X_i) = \sigma^2$. *For large enough n, the distribution of* $\bar{X}$ *will be approximately normal with* $E(\bar{X}) = \mu$ *and* $VAR(\bar{X}) = \frac{\sigma^2}{n}$. *That is,*

$$\bar{X} \approx N(\mu, \frac{\sigma^2}{n}).$$

This theorem it is very important!

Central limit theorem

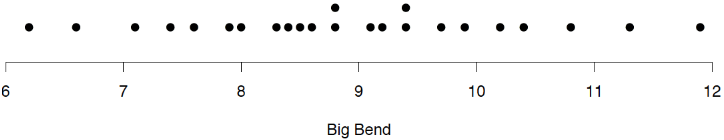
- How large is large enough?

Central limit theorem

- How large is large enough?
- The required size for n depends on the nature of the population distribution of X_i. The closer the distribution of X_i is to normal, the smaller n is required for the approximation to be good.
- For reasonably symmetric distributions with no outliers, $n = 5$ could be sufficient. For distributions with extreme skew or heavy tails/outliers, you may need $n = 100$ or more.
- For much real-world data, $n = 30$ is a relatively safe cut-off, and this sample size is what is typically prescribed to use the CLT.

Outline

Big bend lizards tail length



Big bend lizards tail length

- Our goal is to estimate μ, the population mean tail length in the entire Big bend lizards.

Big bend lizards tail length

- Our goal is to estimate μ, the population mean tail length in the entire Big bend lizards.

```
> bigbend
 [1]  8.8  9.7 10.8  7.1  6.6  9.9 10.2  8.6 10.4 11.9  7.6  8.0  8.5
[16]  7.4  8.3  9.1  9.2  7.9  8.4 11.3  6.2  8.8
> mean(bigbend)
[1] 8.895833
> sd(bigbend)
[1] 1.429953
> length(bigbend)
[1] 24
```

Big bend lizards tail length

- Our goal is to estimate μ, the population mean tail length in the entire Big bend lizards.

```
> bigbend
 [1]  8.8  9.7 10.8  7.1  6.6  9.9 10.2  8.6 10.4 11.9  7.6  8.0  8.5
[16]  7.4  8.3  9.1  9.2  7.9  8.4 11.3  6.2  8.8
> mean(bigbend)
[1] 8.895833
> sd(bigbend)
[1] 1.429953
> length(bigbend)
[1] 24
```

- $\bar{X} = 8.896$ cm is one estimate for μ.

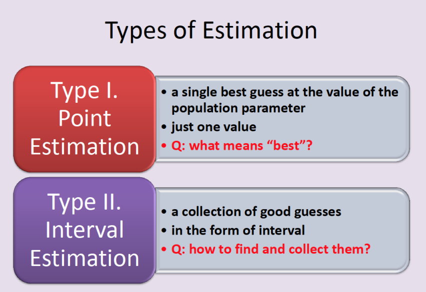
Big bend lizards tail length

- Question: how good is this estimate? How far is μ from 8.896 cm?

Big bend lizards tail length

- Question: how good is this estimate? How far is μ from 8.896 cm?
- The standard error (SE) of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$, but we don't know σ.
- The estimated standard error of $\bar{X}$ is $\frac{S}{\sqrt{n}}$, where S is the sample standard deviation
- S=1.43 and n=24, so estimated SE=$\frac{1.43}{\sqrt{24}} = 0.292$.
- The estimated SE gives us an idea of how far $\bar{X}$ is from μ typically.

The second type of estimation



Outline

1. QQ plot
2. Central limit theorem
3. Review of point estimation
4. The t-distribution
5. Confidence interval

- If $X_1, ..., X_n$ have a normal distribution, then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

The t-distribution

- If $X_1, ..., X_n$ have a normal distribution, then

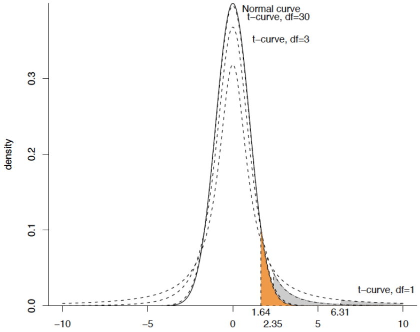
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- When we replace $\sigma/\sqrt{n}$ by estimated SE$=S/\sqrt{n}$,

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim T_v$$

where $v = n - 1$ is called degrees of freedom and T_v is called t-distribution with degrees of freedom v.

The t-distribution



It looks very similar to a standard normal: it's symmetric and bell-shaped, but it is a little more spread out. The amount of additional spread decreases as the degrees of freedom (the sample size) increases.

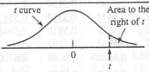
T table

Question: How do we find a value t s.t. $P(T_{10} \geq t) = 0.17$?

T table

Question: How do we find a value t s.t. $P(T_{10} \geq t) = 0.17$?

Table A.8 *t* Curve Tail Areas



t \ ν	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
0.0	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500
0.1	.468	.465	.463	.463	.462.	.462	.462	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461
0.2	.437	.430	.427	.426	.425	.424	.424	.423	.423	.423	.423	.422	.422	.422	.422	.422	.422	.422
0.3	.407	.396	.392	.390	.388	.387	.386	.386	.386	.385	.385	.385	.384	.384	.384	.384	.384	.384
0.4	.379	.364	.358	.355	.353	.352	.351	.350	.349	.349	.348	.348	.348	.347	.347	.347	.347	.347
0.5	.352	.333	.326	.322	.319	.317	.316	.315	.315	.314	.313	.313	.313	.312	.312	.312	.312	.312
0.6	.328	.305	.295	.290	.287	.285	.284	.283	.282	.281	.280	.280	.279	.279	.279	.278	.278	.278
0.7	.306	.278	.267	.261	.258	.255	.253	.252	.251	.250	.249	.249	.248	.247	.247	.247	.247	.246
0.8	.285	.254	.241	.234	.230	.227	.225	.223	.222	.221	.220	.220	.219	.218	.218	.218	.217	.217
0.9	.267	.232	.217	.210	.205	.201	.199	.197	.196	.195	.194	.193	.192	.191	.191	.191	.190	.190
1.0	.250	.211	.196	.187	.182	.178	.175	.173	.172	.170	.169	.169	.168	.167	.167	.166	.166	.165
1.1	.235	.193	.176	.167	.162	.157	.154	.152	.150	.149	.147	.146	.146	.144	.144	.144	.143	.143
1.2	.221	.177	.158	.148	.142	.138	.135	.132	.130	.129	.128	.127	.126	.124	.124	.124	.123	.123
1.3	.209	.162	.142	.132	.125	.121	.117	.115	.113	.111	.110	.109	.108	.107	.107	.106	.105	.105
1.4	.197	.148	.128	.117	.110	.106	.102	.100	.098	.096	.095	.093	.092	.091	.091	.090	.090	.089
1.5	.187	.136	.115	.104	.097	.092	.089	.086	.084	.082	.081	.080	.079	.077	.077	.077	.076	.075
1.6	.178	.125	.104	.092	.085	.080	.077	.074	.072	.070	.069	.068	.067	.065	.065	.065	.064	.064
1.7	.169	.116	.094	.082	.075	.070	.065	.064	.062	.060	.059	.057	.056	.055	.055	.054	.054	.053
1.8	.161	.107	.085	.073	.066	.061	.057	.055	.053	.051	.050	.049	.048	.046	.046	.045	.045	.044
1.9	.154	.099	.077	.065	.058	.053	.050	.047	.045	.043	.042	.041	.040	.038	.038	.038	.037	.037
2.0	.148	.092	.070	.058	.051	.046	.043	.040	.038	.037	.035	.034	.033	.032	.032	.031	.031	.030

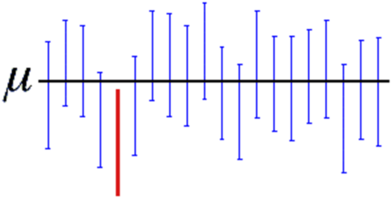
Download T table from our course website.

Outline

Confidence interval for population mean

- Point estimates are almost always wrong.
- Why not collect a lot of good guesses which form an interval, and let the interval cover the population mean with high probability?

Interpretation of a confidence interval



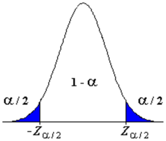
A 95% confidence interval indicates that 19 out of 20 samples (95%) from the same population will produce confidence intervals that contain the population parameter.

In confidence interval, the population mean μ is a fixed unknown constant, the interval is random.

Mechanics of a confidence interval: case 1

If we know the population standard deviation σ,

1. Choose a confidence level $1-\alpha$. Typically, if we require 95% confidence level, then $\alpha = 0.05$.
2. Use z table to find the $z_{\frac{\alpha}{2}}$ critical value such that $P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1-\alpha$.



3. Construct the interval: (L, U), where $L = \bar{X} - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}, U = \bar{X} + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$. (Why do we construct this way?)
4. Conclude: $P(L \leq \mu \leq U) = 1-\alpha$. We are $(1-\alpha) \times 100\%$ confident that the population mean is between (L, U).

Mechanics of a confidence interval: case 2

If we don't know the population standard deviation σ,

1. Choose a confidence level $1 - \alpha$. Typically, if we require 95% confidence level, then $\alpha = 0.05$.
2. Find the value t such that $P(-t \leq T_{n-1} \leq t) = 1 - \alpha$. It also means $P(T_{n-1} \geq t) = \frac{\alpha}{2}$. Use t table with degrees of freedom n-1. We denote the value t as $t_{n-1,\alpha/2}$.
3. Construct the interval: (L, U), where $L = \bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, U = \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$. (Why do we construct this way?)
4. Conclude: $P(L \leq \mu \leq U) = 1 - \alpha$. We are $(1 - \alpha) \times 100\%$ confident that the population mean is between (L, U).

Example: big bend lizards tail length

1. We want to construct a 90% confidence interval for the population mean tail length, so $\alpha = 0.1$.

Example: big bend lizards tail length

1. We want to construct a 90% confidence interval for the population mean tail length, so $\alpha = 0.1$.
2. Find the value t such that $P(T_{23} \geq t) = \frac{\alpha}{2} = 0.05$. Use t table with degrees of freedom n-1=24-1=23.

Example: big bend lizards tail length

1. We want to construct a 90% confidence interval for the population mean tail length, so $\alpha = 0.1$.
2. Find the value t such that $P(T_{23} \geq t) = \frac{\alpha}{2} = 0.05$. Use t table with degrees of freedom n-1=24-1=23.
   t-table gives: t=1.71

Example: big bend lizards tail length

1. We want to construct a 90% confidence interval for the population mean tail length, so $\alpha = 0.1$.
2. Find the value t such that $P(T_{23} \geq t) = \frac{\alpha}{2} = 0.05$. Use t table with degrees of freedom n-1=24-1=23.
   t-table gives: t=1.71
   or use R: qt(0.95, df=23)

Example: big bend lizards tail length

1. We want to construct a 90% confidence interval for the population mean tail length, so $\alpha = 0.1$.
2. Find the value t such that $P(T_{23} \geq t) = \frac{\alpha}{2} = 0.05$. Use t table with degrees of freedom n-1=24-1=23.
   t-table gives: t=1.71
   or use R: qt(0.95, df=23)
3. Construct the interval: (L, U), where

$$L = \bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} = 8.896 - 1.71 * \frac{1.43}{\sqrt{24}} = 8.396,$$

$$U = \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} = 8.896 + 1.71 * \frac{1.43}{\sqrt{24}} = 9.396$$

Example: big bend lizards tail length

1. We want to construct a 90% confidence interval for the population mean tail length, so $\alpha = 0.1$.
2. Find the value t such that $P(T_{23} \geq t) = \frac{\alpha}{2} = 0.05$. Use t table with degrees of freedom n-1=24-1=23.
   t-table gives: t=1.71
   or use R: qt(0.95, df=23)
3. Construct the interval: (L, U), where

$$L = \bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} = 8.896 - 1.71 * \frac{1.43}{\sqrt{24}} = 8.396,$$

$$U = \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} = 8.896 + 1.71 * \frac{1.43}{\sqrt{24}} = 9.396$$

4. Conclude.

Confidence interval R simulation

See R codes from the course webpage.

What's the next?

In the next lecture, we'll discuss sample size and population proportions.