Probability and Statistics
Statistics

Giuliano Casale
Department of Computing, Imperial College London

Maximum Likelihood Estimation

Example: Likelihood for Continuous Distributions

We need to build a reliability model for disk drives in a data center. Let us assume that the time to failure of a disk drive follows an exponential distribution, with density $f(x) = \lambda e^{-\lambda x}$, where x is the time to failure.

We record the following failure times for 10 disks, in hours:
$X = (x_i) = (1200, 1500, 1600, 1700, 1100, 1300, 1400, 1250, 1550, 1650)$
How can we best choose λ ?

Example: Likelihood for Continuous Random Variables

Idea: We seek for the λ value that maximizes the chances of sampling what we observed.
For a continuous distribution, we may do so by choosing λ that maximizes the joint pdf:

$$L(\lambda)=f(X \mid \lambda)=f\left(x_1, \ldots, x_{10} \mid \lambda\right)=\prod_{i=1}^{10} f\left(x_i \mid \lambda\right)=\prod_{i=1}^{10} \lambda e^{-\lambda x_i}$$

where we used independence.
Finally, we use basic calculus to find the λ that maximizes $f(X \mid \lambda)$:

$$\widehat{\lambda}=\arg \max_\lambda L(\lambda)=\frac{10}{\sum_{i=1}^n x_i}=\frac{1}{\bar{X}}=\frac{1}{1425}=0.0007 \text{ failures } / \mathrm{h}$$

where $\bar{X}$ is the observed mean time to failure (MTTF).

Example: Log-Likelihood for Continuous Distributions

To prove the last result, it is easier to work with the logarithm of $L(\lambda)$, which we refer to as the log-likelihood

$$\ell(\lambda)=\log L(\lambda)=\log \prod_{i=1}^n \lambda e^{-\lambda x_i}=n \log \lambda-\lambda \sum_{i=1}^n x_i$$

Since the logarithm is monotone, we can equivalently maximize $\ell(\lambda)$ instead of $L(\lambda)$. In this case, the maximum is found by setting

$$\ell^{\prime}(\lambda)=\frac{d}{d \lambda} \ell(\lambda)=\frac{n}{\lambda}-\sum_{i=1}^n x_i=0$$

that yields $\widehat{\lambda}=\frac{n}{\sum_{i=1}^n x_i}=\frac{1}{\bar{x}}$. Since $\ell^{\prime \prime}(\widehat{\lambda})<0$, this is a maximum.

Example: Likelihood for Discrete Distributions

- Let us now assume a discrete setting, where we count every day the number of disks x that fail. The data center has $m=20000$ disks.

- Over the last $d=100$ days, we observe the following data:

x	0	1	2	3	4	5	6	7	8	9	10
Frequency	2	16	35	22	21	3	1	0	0	0	0

- Assuming that disks are independent, we treat our dataset as a sample of Binomial(m, θ) with $p(x \mid \theta)=\binom{m}{x} \theta^x(1-\theta)^{m-x}$, where θ is the (daily) disk failure probability.

- How shall we estimate θ ?

Example: Likelihood for Discrete Distributions

Idea: again, we seek for the θ value that maximizes the chances of sampling what we observed.
For a discrete distribution, we model this using the joint pmf:

$$L(\theta)=p\left(x_1,\ldots,x_d \mid \theta\right)=\prod_{i=1}^{d}\binom{m}{x_i}\theta^{x_i}(1-\theta)^{m-x_i}$$

where the last passage follows again from independence.
Lastly, we find the θ that maximizes $L(\theta)$, which is

$$\widehat{\theta}=\arg\max_{\theta} L(\theta)=\frac{1}{d}\frac{\sum_{i=1}^{d} x_i}{m}=0.0001285 \text{ failure probability}$$

(Left to check as an exercise.)

Maximum Likelihood Estimation

(1) The likelihood function, $L(\theta)=\prod_{i=1}^{n} f\left(x_i \mid \theta\right)$ is the product of the $n$pmf/pdf viewed as a function of a parameter θ.
(2) Take the natural log of the likelihood to get the log-likelihood function $\ell(\hat{\theta})=\log(L(\theta))$ and collect terms involving θ.
(3) Find the value of θ for which log-likelihood is maximised. This is typically done by finding $\hat{\theta}$ that solves

$$\ell'(\hat{\theta})=\frac{d}{d\theta}\log(L(\hat{\theta}))=0$$

(1) If the estimate $\hat{\theta}$ obtained in step 3 corresponds to a maximum $\frac{d^2}{d\theta^2}\ell(\hat{\theta})<0$, then $\hat{\theta}$ is confirmed as the maximum likelihood estimator (MLE) of θ.

Further remarks on MLE

- In large sample sizes, the MLE progressively becomes unbiased, efficient and consistent. This can be proved under mild technical assumptions.
- In small sample sizes there is no such guarantee and the quality of a MLE can vary. Bayesian parameter estimation is an area of statistics that deals with this problem.
- For single parameter MLE, if θ is discrete and $\ell(\cdot)$ unimodal, then we can run MLE as usual, compare $\lceil\hat{\theta}\rceil$ and $\lfloor\hat{\theta}\rfloor$, and return the one having the largest likelihood. Otherwise, more complex algorithms are required to search for the MLE.

- MLE generalizes to multi-parameter distributions. Yet this requires multivariate calculus and the maximization may give more than one answer if ℓ has several peaks (local maxima).

An alternative to MLE: the method of moments

- In some cases, we may only know statistics of a distribution, such as its sample mean or sample variance.
- In this setting, MLE is not viable but we can still estimate parameters if we have enough sample moments.
- This approach, called moment matching (or method of moments), tries to match the true and sample moments.
- For example, in the disk drive reliability model we had

$$\lambda = \frac{1}{E[X]} \quad \bar{X} = 1425$$

and matching $E[X] = \bar{X}$ yields $\lambda = 1/\bar{X} = 1/1425 = 0.0007$.

- This is effective on simple models, but unlike MLE can suffer biases for more complex distributions.

Central Limit Theorem

The Central Limit Theorem (CLT)

The CLT is a general result for sums of random variables. In statistics, it helps to study the distribution of the sample mean.

Let $X_1, X_2, \ldots, X_n$ be now n independent and identically distributed (i.i.d.) random variables from any probability distribution with mean μ and variance σ^2 both finite.

- We know that $\mathrm{E}\left(S_n\right) = n\mu$ and $\mathrm{Var}\left(S_n\right) = n\sigma^2$.
- Thus, we have $\mathrm{E}\left(S_n - n\mu\right) = 0$ and $\mathrm{Var}\left(S_n - n\mu\right) = n\sigma^2$.
- Dividing by $\sqrt{n}\sigma$,

$$\mathrm{E}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma}\right) = 0, \quad \mathrm{Var}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma}\right) = 1$$

However, what can we say about the underlying distribution?

We have the following celebrated result:

Central Limit Theorem (CLT)

$$\lim_{n\rightarrow\infty} \frac{S_n - n\mu}{\sqrt{n}\sigma} \sim \mathrm{N}(0,1)$$

This result can also be written as

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

where $\bar{X} = \frac{S_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$ is the sample mean.

Implications for the Sample Mean

The CLT thus implies that for large, but finite, n

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

- A rule of thumb for "large n " is often $n \geq 30$.
- Amazingly, this result holds irrespective of the distribution of the $\{X_i\}$ (and including discrete random variables). CLT thus demonstrates that statistical regularity can arise even from the combination of highly-diverse random phenomena.
- If $X_i \sim \mathrm{N}\left(\mu, \sigma^2\right), \forall i$, the result becomes exact even for finite n, since the sum of independent normal random variables is normally distributed (proof via mgfs).

Sketch proof of the CLT

Given the n i.i.d. r.vs. $X_1, X_2, \ldots, X_n$, standardise their sum to get

$$Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} = \frac{\sum_{i=1}^{n}\left(X_i - \mu\right)}{\sqrt{n}\sigma} = \sum_{i=1}^{n} \frac{Y_i}{\sqrt{n}\sigma}$$

where $Y_i = X_i - \mu$ is a shifted version of X_i with zero mean.
The moment generating function (mgf) of Z_n is

$$M_{Z_n}(t) = \left(M_Y\left(\frac{t}{\sqrt{n}\sigma}\right)\right)^n$$

where M_Y is the mgf of Y_1 (and each Y_i).

- We can now expand $M_Y(t)$ at zero using Taylor's theorem

$$M_Y(t)=M_Y(0)+M_Y'(0)t+\frac{1}{2}M_Y''(0)t^2+O\left(t^3\right)=1+\frac{1}{2}\sigma^2t^2+O\left(t^3\right)$$

where we note that the derivatives of the mgf are:
$M_Y'(0)=E\left(Y_i\right)=0$ and
$M_Y''(0)=E\left(Y_i^2\right)=\sigma^2+E\left(Y_i\right)^2=\sigma^2$.

- Plugging the result back into $M_{Z_n}(t)$, we can see that

$$\lim_{n\to+\infty}M_{Z_n}(t)=\lim_{n\to+\infty}\left(1+\frac{t^2}{2n}+O\left(n^{-3/2}\right)\right)^n\to e^{t^2/2}$$

- But $e^{t^2/2}$ is the mgf of the standard Normal distribution.

Application to coin tossing

Consider the simplest example, where $X_1, X_2, \ldots$ are i.i.d. Bernoulli (p) discrete random variables taking values 0 or 1 .

- Then the $\{X_i\}$ have mean $\mu=p$ and variance $\sigma^2=p(1-p)$.
- Thus, by definition, for any n,

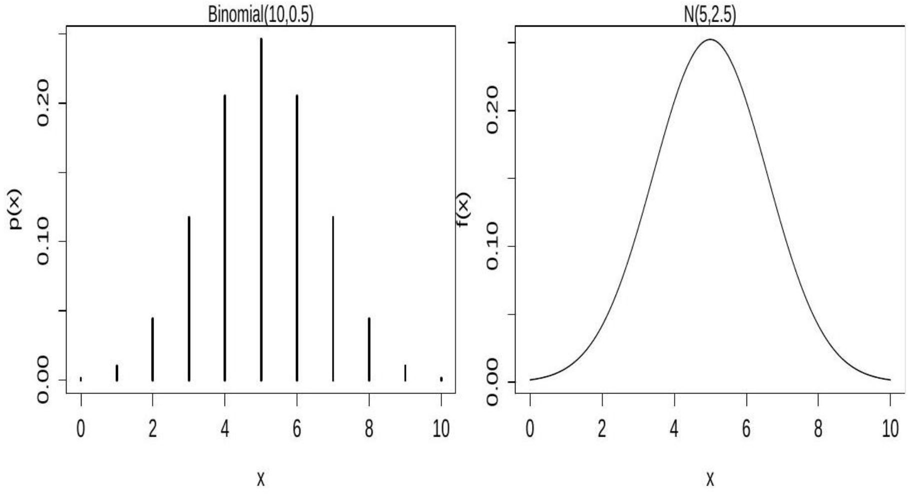
$$\sum_{i=1}^n X_i\sim\text{Binomial}(n,p)$$

which has mean np and variance $np(1-p)$.

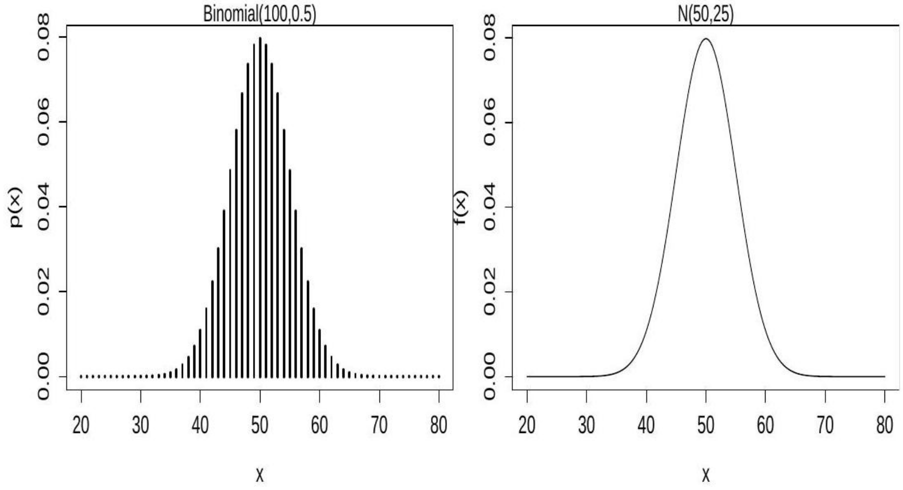
- But now, by the CLT

$$\text{Binomial}(n,p)\approx\mathrm{N}\left(n\mu,n\sigma^2\right)\equiv\mathrm{N}(np,np(1-p))$$

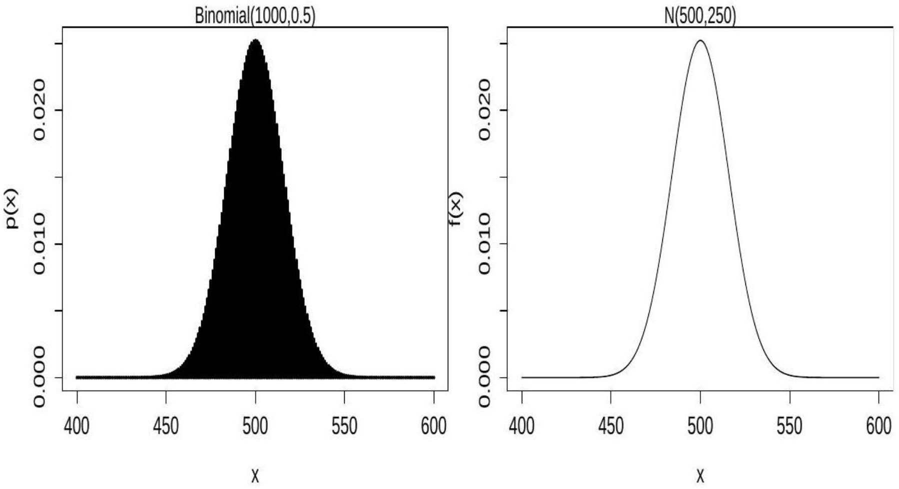
Binomial $\left(10, \frac{1}{2}\right)$ pmf & N(5,2.5) pdf



Binomial $\left(100, \frac{1}{2}\right)$ pmf & N $(50, 25)$ pdf



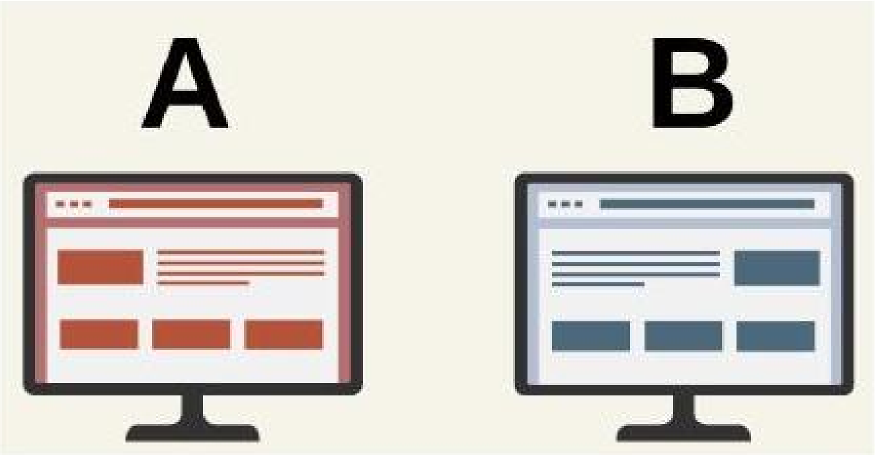
Binomial $\left(1000, \frac{1}{2}\right)$ pmf & N $(500, 250)$ pdf



Hypothesis Testing

Motivation

Suppose a video streaming company has developed a new feature for their tv app that they believe will increase user engagement. Before rolling it out to all users, they decide to conduct a so-called A/B test.



A group of users (control group) will continue to use the current version of the app without the new feature, while another group (test group) will use the version of the app with the new feature.

Hypothesis Testing

- Consider the following two hypotheses:
- The Null Hypothesis: there is no difference in user engagement between the two groups.
- The Alternative Hypothesis: a difference exists, e.g., the new feature increases user engagement.
- If the test group shows much higher engagement than the control group, then we will reject the null hypothesis and conclude that the evidence favours the alternative hypothesis.
- This is an example of hypothesis testing. The idea is to assess the strength of the evidence in the sample data by which we can reject (or retain) the null hypothesis.

Parametric tests

- Suppose again we have a random i.i.d. sample $(X_1, \ldots, X_n)$ of a random variable X from an unknown distribution P.
- Parametric tests typically assume the sample comes from a parametric family $P(\cdot \mid \theta)$ and test whether we could reasonably assume $\theta = \theta_0$ for some particular value θ_0.
- In the A/B testing example, we may set X as the difference in user engagement feedbacks in the two groups and test if $\mu = 0$ under the assumption that $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$.

Hypotheses

- Formally, let H_0 be the null hypothesis and H_1 be the alternative hypothesis.
- Most often we simply test

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

This is known as a two-sided test.

- A one-sided test is also possible, e.g.:

$$\begin{array}{lll} H_0 : \theta = \theta_0 & \text{versus} & H_1 : \theta < \theta_0 \\ H_0 : \theta = \theta_0 & \text{versus} & H_1 : \theta > \theta_0 \end{array}$$

- Usually, the null hypothesis is formulated with an equality sign $(=)$, while the alternative hypothesis uses one of $(\neq, <, >)$.

Rejection Region for a Test Statistic

- To test the validity of H_0, we choose a test statistic $T(X)$ of the data for which we can find the distribution under H_0.
- The "art" of hypothesis testing is to define the test by identifying a rejection region $R \subseteq \mathbb{R}$ of low probability values of T under the assumption that H_0 is true, so that

$$P\left(T \in R \mid H_0\right)=\alpha$$

for some small probability α (say 5%). We call α the significance level of the test.

- Rule: A well chosen rejection region will have relatively high probability under H_1, whilst retaining low probability under H_0.
- We calculate the observed test statistic $t(x)$ for our sample x :
  (1) If $t \in R$ we "reject the null hypothesis at the $100\alpha\%$ level".
  (2) If $t \notin R$ we "retain the null hypothesis at the $100\alpha\%$ level".

Testing for Population Mean - Known Variance

- Suppose $X_1, \ldots, X_n$ are i.i.d. $N\left(\mu, \sigma^2\right)$ with only σ^2 known.
- We may wish to test if $\mu=\mu_0$ for some specific value μ_0.
- Then we can state our null and alternative hypotheses as

$$H_0: \mu=\mu_0 \text { versus } H_1: \mu \neq \mu_0$$

- Under $H_0: \mu=\mu_0$, we then know both μ and σ^2. So for the sample mean $\bar{X}$ we know from the CLT that

$$Z=\frac{\bar{X}-\mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$$

where $\sigma / \sqrt{n}$ is often called in statistics the standard error.

- By the CLT, the result also holds approximately when the X_i are not normally distributed.

Testing for Population Mean - Known Variance

- If the Z-test statistic takes "extreme" values far from zero, there is evidence to conclude that H_0 should be rejected. Otherwise, the data is inconclusive and we need to retain H_0.
- So we may define our rejection region R to be the $100\alpha\%$ tails of the standard normal distribution distribution,

$$R = \left(-\infty, -z_{1-\frac{\alpha}{2}}\right) \cup \left(z_{1-\frac{\alpha}{2}}, \infty\right)$$

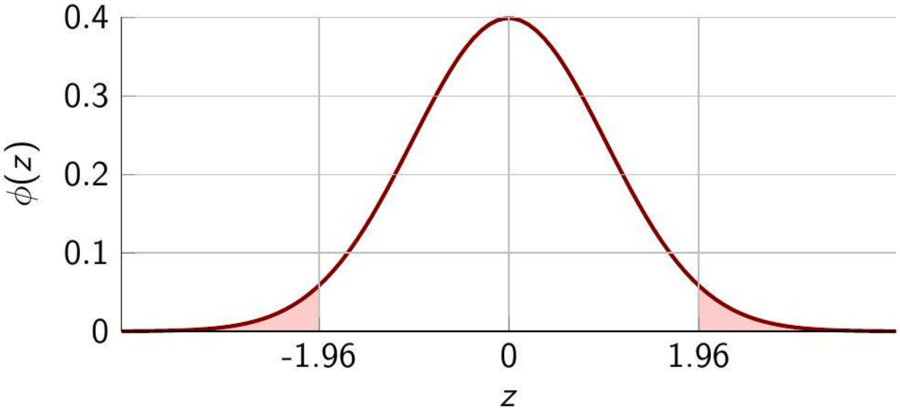
we have $P\left(Z \in R \mid H_0\right) = \alpha$.

- We thus reject H_0 at the $100\alpha\%$ significance level exactly when our observed test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \in R$$

Testing for Population Mean - Known Variance

- Rejection region R for $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ at the 5% level:



- Each tail includes $\frac{\alpha}{2} = 2.5\%$ of the mass under $\phi(z)$.

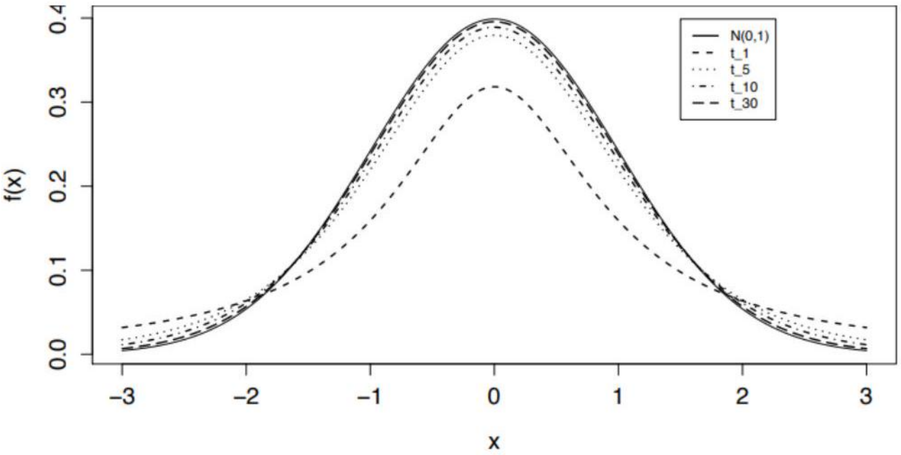
Dealing with unknown variance

- As we measure a population (simulated or real), we may not know the real variance σ^2, but only the (bias-corrected) sample variance S^2. How do the formulas change in this case?

- The problem was studied by William S. Gosset, who published a paper in 1908 under the pseudonym Student.
- The study introduced the Student's t distribution, a generalization of the Normal distribution, widely used in particular for statistical analysis of small samples.
- This is the basis of the **t**-test statistic.

Student's t distribution

t_ν : Student's t distribution with ν degrees of freedom.



Student's t distribution table

Statistical table for the p-quantiles $t_{\nu,p}$ of the Student's t distribution with ν degrees of freedom:

	quantile					quantile			
ν	0.95	0.975	0.99	0.995	ν	0.95	0.975	0.99	0.995
1	6.31	12.71	31.82	63.66	9	1.83	2.26	2.82	3.25
2	2.92	4.30	6.96	9.92	10	1.81	2.23	2.76	3.17
3	2.35	3.18	4.54	5.84	12	1.78	2.18	2.68	3.05
4	2.13	2.78	3.75	4.60	15	1.75	2.13	2.60	2.95
5	2.02	2.57	3.36	4.03	20	1.72	2.09	2.53	2.85
6	1.94	2.45	3.14	3.71	25	1.71	2.06	2.48	2.78
7	1.89	2.36	3.00	3.50	40	1.68	2.02	2.42	2.70
8	1.86	2.31	2.90	3.36	∞	1.645	1.96	2.326	2.576

Testing for Population Mean - Unknown Variance

- If σ^2 in the previous example were unknown, we would have

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

which uses the Student's t distribution with $n-1$ degrees of freedom (t_{n-1}) and the bias-corrected sample std. dev. S.

- So for a test of

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu \neq \mu_0$$

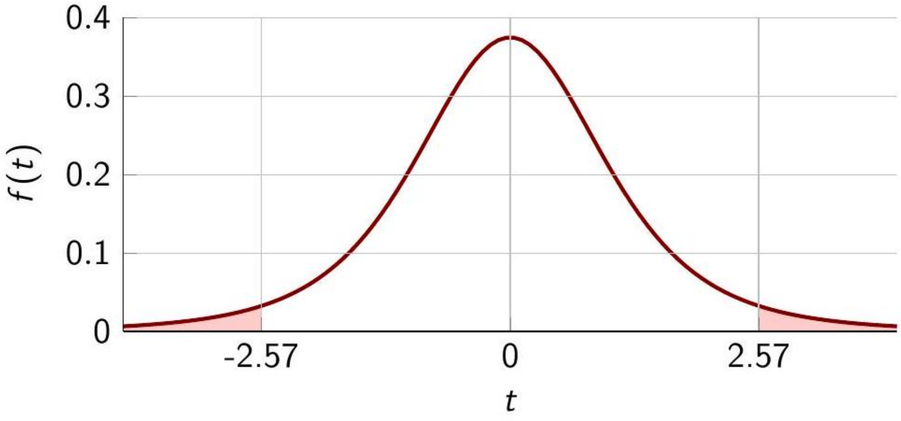
at the α level, the rejection region of our observed test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

is then given by $R = \left(-\infty, -t_{n-1,1-\frac{\alpha}{2}}\right) \cup \left(t_{n-1,1-\frac{\alpha}{2}}, \infty\right)$

Testing for Population Mean - Unknown Variance

- Rejection region R for $n = 6$ and $t = \frac{\bar{x}-\mu_0}{s/\sqrt{6}}$ at the 5% level:



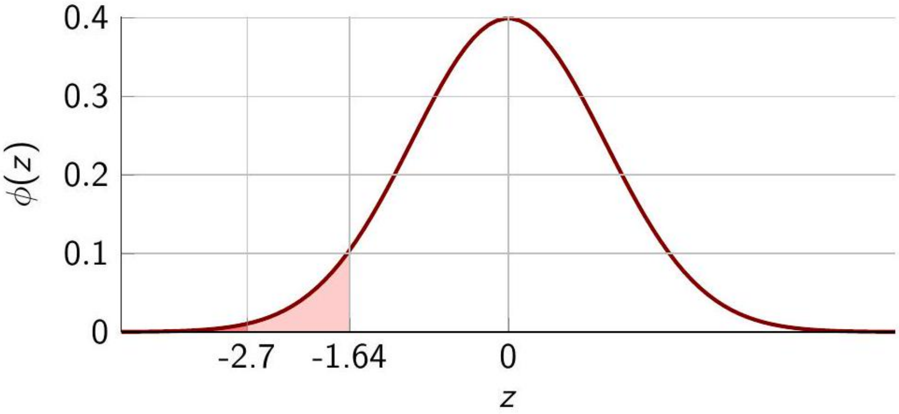
p-Values

- Often, it is important to quantify the statistical significance of a result, in addition to giving a reject/retain outcome.
- The **p**-value of the data is the probability of obtaining a test statistic at least as extreme as the one actually observed, assuming H_0 is correct.

- In other words, the p -value is the maximum significance level at which we still reject the null hypothesis H_0 for that sample.
- Thus, if we are given a fixed α, the null hypothesis H_0 is rejected if the p-value is less than or equal to α.
- Rule: Smaller p-values suggest stronger evidence against H_0.

p-Value in a One-sided Lower-Tailed Test ($H_1 : \theta < \theta_0$)

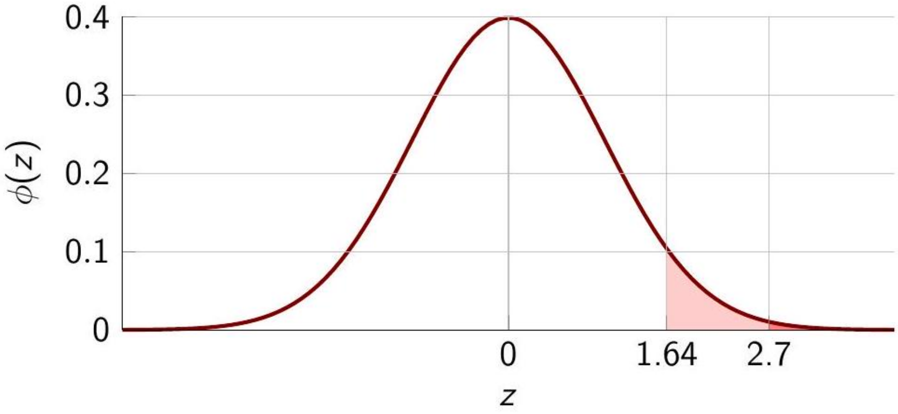
- With known variance, the p-value is $\Phi(z)$.
- With unknown variance, the p-value is $F(t)$, where $F(\cdot)$ is the cdf of the Student's t distribution.



$z = -2.7$ and $H_1 : \theta < \theta_0$ then $p = 0.003467$. The lighter shading is R for $\alpha = 0.05$.

p-Value in a One-sided Upper-Tailed Test ($H_1 : \theta > \theta_0$)

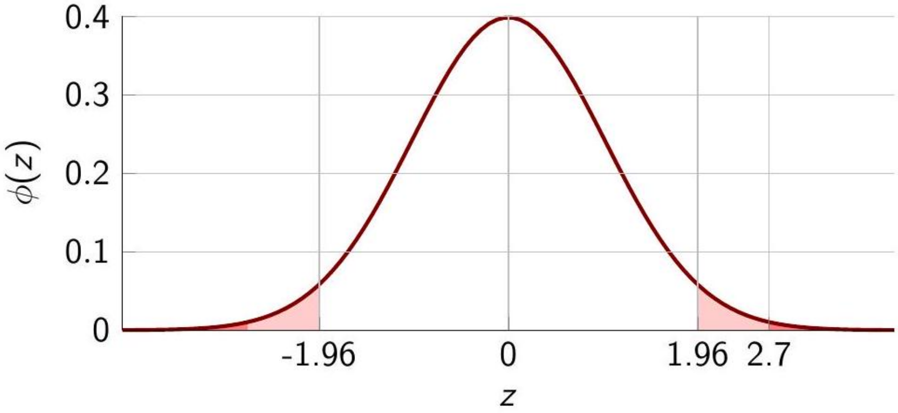
- With known variance, the p-value is $1 - \Phi(z)$.
- With unknown variance, the p-value is $1 - F(t)$.



If $z = 2.7$ and $H_1 : \theta > \theta_0$ then $p = 0.003467$. The lighter shading is R for $\alpha = 0.05$.

p-Value in a Two-sided Test

- With known variance, the p-value is $2 \times (1 - \Phi(|z|))$.
- With unknown variance, the p-value is $2 \times (1 - F(|t|))$.



For example, if $z = 2.7$ then the two-sided **p**-value is $p = 0.006934$. The lighter shading is R for $\alpha = 0.05$.

Other commonly used tests (Unassessed)

Besides the Z-test and the t-test, several other tests exist, e.g.:

- Paired samples tests: Extensions exists of Z-test and t-test to compare samples in the form $(X_1, X_1'), \ldots, (X_n, X_n')$.

- Chi-squared test: Used to assess whether observed frequencies of a random variable differ from expected.

- Kolmogorov-Smirnov test: Used to determine if two samples come from the same distribution or to compare a sample with a reference probability distribution.